Norwegian School of Economics

Bergen, Fall 2022

Predicting Private Equity Fund Performance with Machine Learning

Nikita Kruglikov & Andreas Forthun

Supervisor: Håkon Otneim

Master thesis, Economics and Business Administration

Major: Financial Economics, Business Analytics

NORWEGIAN SCHOOL OF ECONOMICS

Abstract

This paper has the objective of applying machine learning models to predict the performance of private equity funds, to allow for more effective fund selection for investors in the private markets. Prior research has mainly focused on determining a probability of private equity funds exceeding a pre-defined rate of return, or on examining factors which influence the returns of said funds. We instead utilize the factors previously determined to influence private equity fund returns to train machine learning algorithms predicting the returns investors can expect to receive from the moment of making a primary investment into the fund, until the fund's liquidation. Due to it being the measure of choice for both general partners (GPs) and limited partners (LPs) in the private equity industry, we selected the Net Internal Rate of Return (NIRR) as our measure of return. We mainly source our data from PitchBook, which allows us to form a more extensive set of predictor variables, while supplementing this data with macroeconomic variables collected from public sources. To estimate predictor models, we apply machine learning methodologies including stepwise regression methods, such as the Akaike Information Criterion and Ridge, as well as more advanced methods consisting of Support Vector Machine and Bayesian Regularized Neural Networks. The latter enables us to add flexibility into our models by considering interaction effects between predictor variables. Our models show favorable results, with the Support Vector Machine giving the strongest performance on in-sample data, delivering a mean squared error (MSE) value of 0.0072. This does however come at the expense of weaker performance on the out-of-sample data, with the model achieving an MSE of 0.0538 on the test set, likely implying that the model overfits the data when calculating the algorithm for the training set. This is compensated by the linear Akaike Information Criterion model performing quite strongly on the out-of-sample, displaying an MSE value of 0.0370.

Table of Contents

1. Introduction 4

2. Related Literature 9

3. Theory 10

3.1 Private Equity 10

3.2 Machine Learning 16

4. Data 18

4.1 Data Selection 18

4.2 Data Transformation 26

4.3 Dataset Overview 29

5. Methodology 32

5.1 Test-Train and Overfitting 32

5.2 Ordinary Least Squares 33

5.3 Stepwise Regression 34

5.4 Ridge Regression 36

5.5 Support Vector Regression 37

5.6 Artificial Neural Networks 40

6. Results 42

6.1 Models' Precision 42

6.2 Analysis of Coefficients 44

7. Conclusion 55

7.1 Conclusion 55

7.2 Biases, Limitations & Future Research 56

Bibliography 59

Data Sources 63

Appendices 65

Terminology

PE – Private Equity

GP – General Partner

LP – Limited Partner

PPM – Private Placement Memorandum

NAV – Net Asset Value

NIRR – Net Internal Rate of Return

TVPI – Total Value to Paid-In (Capital)

PME – Public Market Equivalent

CF – Cash Flow

MUSD – Million United States Dollars

ML – Machine Learning

NN – Neural Network

SVM – Support Vector Machine

MSE – Mean Squared Error

OVR – One-Versus-Rest

LDA – Linear Discriminant Analysis

SVC – Support Vector Classification

AUC – Area Under the (ROC) Curve

OLS – Ordinary Least Squares

SEC – U.S. Securities and Exchange Commission

ESG – Environmental, Social, and Governance

M&A – Mergers & Acquisitions

GICS – Global Industry Classification Standard

LOOCV – Leave-One-Out Cross-Validation

SVR – Support Vector Regression

RBF – Radial Basis Function

1. Introduction

Research shows that private equity achieves higher returns and lower volatility than the S&P 500, an often-used benchmark of public market performance (Cambridge Associate LLC, 2022). Additionally, private capital median net IRRs exceeded 10% on average, making private equity an attractive asset class, as illustrated in Figure 1. Due to a decline in the number of IPOs (Gupta, Koller, & Stumpner, 2021), an extended period of low interest rates, a decline in the number of public firms, a rise in unicorn firms that stay private longer, and other geo-political considerations, the public markets have become less attractive relative to the private market. However, private markets have historically been a domain solely for institutional investors on the background of the industry's restrictive entry requirements. A minimum requirement of a $5 million initial investment has been a requirement to be accepted as a Qualified Purchaser, which effectively excluded non-institutional investors from the private market (Investment Company Act of 1940, 1940).

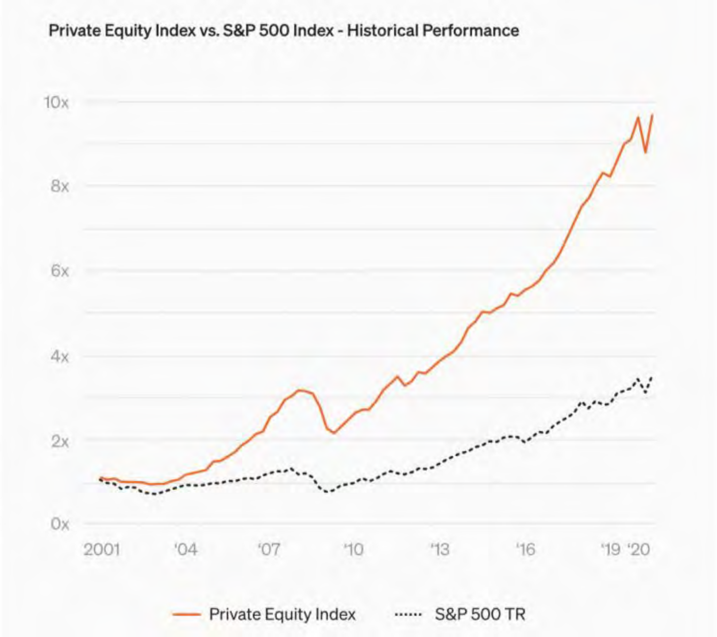


Figure 1: PE returns compared to S&P500 returns (Moonfare GmbH, 2022)

The SEC is however exploring ways to democratize private markets. In 2020, they amended the definition of "Accredited Investor" and expanded the definition of "Qualified Institutional Buyer" in the Securities Act of 1933 (Securities and Exchange Commission, 2020). This change meant that individuals with either a minimum annual income of $200,000 or net worth of $1 million could gain accreditation. The new definition covers around 13.6 million US households, 1.5 million Qualified Purchaser households, and 75% of total wealth in America. In June 2020, the U.S. department of Labor published an information letter clarifying that under the Employee Retirement Income Security Act of 1974 fiduciaries of 401(k) plans are allowed to include diversified investment options with private equity exposure if certain requirements are met (Paul Hastings, 2020). The Investment Company Act of 1940 aims to achieve more stringent reporting requirements, lower investment minimums, leverage limits, simpler tax reporting requirements, and the creation of independent boards designed to represent shareholders' interests.

Due to the composition of private markets, they can provide opportunities for excess or uncorrelated returns. A decline in the number of listed companies (U.S.) from 8,090 in 1996 to 4,266 in 2019 have made true diversification without access to the private market increasingly difficult (The World Bank, 2022). There are currently 1,191 privately held start-ups with valuations of over $1 billion (unicorns) and a total valuation of $3,682 billion (CB Insights, 2022). Additionally, portfolios and institutional investors allocate on average around 15-20% of their investments to private equity and top-decile performers often have 40% or more of their portfolio in private markets, suggesting that an allocation to private markets is necessary for successful portfolio management (Sexton & Veronis, 2021).

Despite the positive prospects, the actual return of a private equity fund is difficult to measure accurately (McSwain, 2019). Returns are often presented as "net to limited partner", but with a disclosure that "due to the fundamental differences between [how private equity and public market returns are calculated], direct comparison … is not recommended." For a more accurate means of comparing private investment performance relative to public alternatives, "investors should look to adjusted public market returns." Furthermore, "the timing and magnitude of fund cash flows are integral to the … performance calculation." Thus, funds can use "fund level engineering" that may "optically boost" the limited partner returns by 3% or more. This boosting implies that private equity funds advertise returns that 'over some periods of time', 'no client received'.

Critics have also expressed concern regarding the democratization of the private market in that this might "worsen wealth inequality by sucking a huge pile of money out of the pockets of workers saving for retirement and shepherding it to the few fabulously wealthy owners of private equity firms" and about investment options that are too risky or complex for retail investors to evaluate (Alon-Beck, 2020). Howard Marks, CFA, noted that there is no easy way to evaluate private investment returns and so "complex, multi-dimensional analysis is required." With higher fees, lower liquidity, and less regulatory oversight retail investors would not have the same protection and available information as they do in public markets, while having increased risk. A report on motivation among managers show 28% report staying in the investment industry to help clients and 36% believing that acting in their clients' best interests implies taking on career risk.

SEC Office of Compliance Inspections and Examinations issued a Risk alert in June 2020 about problems identified when examining enforcement actions involving private fund advisers (Office of Compliance Inspections and Examinations, 2020). This alert addresses a lack of disclosures of potential conflicts of interest, excessive fees charged, and a lack of policies and procedures regarding codes of ethics and insider trading.

First-generation retail-investor focused private equity funds incurred problems with inconsistent performance, liquidity mismatch issues, higher fees and being less aligned with shareholders (Sexton & Veronis, 2021). Distributions that were paid out came out of offering proceeds, leverage was utilized improperly, and the underlying beta coefficient was hidden behind the notion of a fund being "non-traded". However, larger private equity investment managers are increasingly entering the investor-focused private market and have brought with them institutional style pricing, structure, transparency, and information. By providing access to experienced managers, the SEC could enhance protection of retail investors by introducing a "scale and experience" criteria for managers (Comittee on Capital Markets Regulation, 2019). Additionally, limiting retail access to managers with an institutional investor base may help ensure that investors are exposed to experienced private markets managers only. For example, requiring regulated funds of private funds to only invest in private funds that accepts more than a certain percentage of their capital commitments from institutional investors. This is to the benefit of retail investors, as institutional investors negotiate favorable terms for their investments and provides an opportunity for retail investors to achieve incentive alignment with institutional investors.

Commentators recommends embracing regulated funds, as their legal structure provides core investor protections and regulated funds are managed by registered investment advisers who owe a fiduciary duty to the fund through being subject to oversight of an independent board, as well as distributing investments and their payoffs through intermediaries who must act in the best interests of the investors. A regulated fund also designs a diversified portfolio with the intention of reducing the risk that losses at any one underlying company will outweigh successful investments. These funds also have substantial resources and sophisticated routines for reviewing and performing due diligence of investments on behalf of the fund's investors.

Private investment funds should be interested in tapping the individual investor market. An increase in global individual investable assets is also expected from $70T in 2018 to $106T in 2025 (Finley, 2019). With an average allocation of less than 5% to alternative investment among individual investors, compared to almost 30% for pension funds and individual investors, there is certainly potential for further allocation to private markets with increased access for retail investors.

By creating an easy-to-apply algorithm that only utilizes variables with a proven correlation to the net internal rate of return and total value paid-in, retail investors can make informed investment choices in private markets. The advantages of this include higher average returns to investors and a larger private equity market with a better structure for allocating capital to sound ideas and companies. By defining a linear model and using stepwise selection, we clearly define the variables which are relevant for investors looking for sound investments in the private equity market. Furthermore, we use machine learning techniques to build prediction models. These include support vector regression and neural networks. Although these methods have more flexibility and are thus better able to model non-linear relationships between the predictors and the response variables, they are also dependent on large datasets. Additionally, such flexible models are also more prone to overfitting, which we also find in our results. While the support vector regression model achieves a much lower MSE on in-sample data, it also has a marked decline when used on out-of-sample data.

Thesis Outline

This thesis will include seven sections. In Section 2 (Related Literature), we describe past literature where machine learning methods have been used as a tool for predicting private equity performance. Further, in Section 3 (Theory), and more specifically in Section 3.1 (Private Equity), the details regarding private equity which are necessary for understanding this paper are described. In the following section, 3.2 (Machine Learning), a broad overview of the field of Machine Learning is presented. Section 4 (Data) contains three sections which describe: the process of selecting relevant data, 4.1 (Data Selection), the process of transforming this data into a suitable dataset, 4.2 (Data Transformation), and the final dataset overview, 4.3 (Dataset Overview). In Section 5 (Methodology), we go further into details regarding the approach we have used for building prediction models. Firstly, theory regarding the statistical phenomenon of overfitting is presented in 5.1 (Test-Train and Overfitting). Next, the methods for building prediction models are presented in 5.2 (Ordinary Least Squares), 5.3 (Ridge Regression), 5.4 (Support Vector Regression), and finally 5.5 (Artificial Neural Networks). In Section 6 (Results), a two-part analysis of the results is presented as 5.1 (Models' precision) where we compare the prediction accuracy of the models and in 5.2 (Analysis of Coefficients), where we take a detailed look at the coefficients. Finally, in Section 7 (Conclusion), and more specifically in 7.1 (Conclusion), we present a summary of this paper, and in 7.2 (Biases, Limitations & Further Research), we describe biases and limitations that may have affected the results in this paper as well as proposals for further research.

2. Related Literature

This section will provide an overview of previous literature concerning the use of machine learning in predicting private equity performance. As computational power is becoming increasingly commercially available, several papers have been written recently with the aim of predicting performance in private equity. Since the processing power needed to estimate complex prediction models has not been available for long, related literature is relatively new and leaves many fields to be explored and discoveries to be refined.

In 2019, a master thesis was written about machine learning- and survival prediction models for decision support in the private equity market (Tiozzo & Morales, 2019). Using random forest and neural network models, in both the multilayer perceptron and long short-term memory configurations, the authors predict the probability associated to the future state of a private company. These states are involve being acquired, going bankrupt, staying private or going public. Both the neural networks and random forest models provided strong results in terms of predictive performance. The tuned random forest model reaches a high accuracy on each one-versus-rest (OVR) classification and very high minority class recall.

This paper suggests that there are significant relationships identifiable by machine learning models, which can be used to reliably predict a private equity fund performance. An unpublished paper by Pachnanda and Raj further confirms this, by using machine learning to predict a constructed target variable that assumes the value of 1 if the predicted public market equivalent (PME) is above 1 and 0, otherwise. The PME used in this case is the measure defined by Kaplan & Schoar, which equals the ratio of total distributions over contributions, discounted with a public market index. Their best models; Logistic, Linear Discriminant Analysis (LDA), and Support Vector Classification (SVC), achieve an accuracy of 69 % and an Area under the ROC Curve (AUC) of 0.56.

An article by asset management firm Unigestion shows how professionals investors are beginning to use machine learning models to complement the due diligence process and achieve higher returns (Sigrist & Perfetto, 2019). Unigestion uses Ordinary least squares (OLS), gradient boosting and random forest models as complements to human judgment. The models are designed to determine the probability that a fund's performance will exceed a pre-defined hurdle rate. These papers focus on classification, however, few of them attempt to perform a quantitative prediction of returns at liquidation with variables known at close. This is the area of research we are attempting to pioneer and hopefully spur further research in.

3. Theory

3.1 Private Equity

Private equity can broadly be defined as the universe of investments into non-publicly trading companies by private equity houses specialising in such investments (Lerner, Leamon, & Hardymon, 2012). Private equity houses usually have three main investment types – Buyouts, Venture, and Growth Equity. Buyouts involve purchasing a majority stake in mature companies which generate profit, and using the majority stake for active management and operational improvement of companies, in order to later exit these companies for a profit. Buyout funds often finance their acquisition of companies with a portion of debt, giving rise to the term "leveraged buyouts". Venture on the other hand, focuses on minority stakes in younger companies which are not yet profitable but have high growth potential. As this involves a higher risk, there is less leverage used in Venture. Finally, Growth Equity combines the two previous private equity types by focusing on minority stakes in mature companies which need capital for a transformation of the company with potential for high growth.

3.1.1 Fund Structure

Private equity funds are structured as financial vehicles with a limited lifetime which usually last for 10-15 years (Lerner, Leamon, & Hardymon, 2012). Two main players usually take part in a private equity fund – the general partner (GP) and the limited partners (LPs). The general partner is the management of the fund, comprised of partners from the private equity firm which launches the fund, and these have the task of finding and acquiring promising companies, managing these investments and operationally improving the companies, and later exiting the investments for a higher value than they were bought for, generating profit. Meanwhile, as general partners need capital to conduct their investments, limited partners include investors who provide capital to the general partners' funds. Limited partners do not involve themselves in the active management of a fund, as this is the duty of the general partner. Potential investors usually receive a Private Placement Memorandum (PPM) from the general partner, describing the fund's strategy of how the fund plans to generate profit, the experience of the management team, and potentially previous funds the management team has launched and their performance.

General partners are compensated for their work in managing the limited partners' capital through two ways. Firstly, similarly to mutual funds, private equity fund managers receive a fixed management fee annually, which is usually set at 1.5-2.5% of the amount the limited partners have committed to the fund, i.e., have agreed to invest in the fund. Additionally, private equity fund managers receive a variable compensation – "carried interest", which is defined as a percentage of the fund's profits received by the general partner after the limited partners receive their initial investment amount back. The carried interest is often set at 20% of profits after the limited partners' initial investment amount is returned. Such non-equal distribution of returns through the variable carried interest incentivizes the general partner to work most efficiently with managing the fund and its portfolio companies. The structure of a private equity fund is summarized in Figure 2.

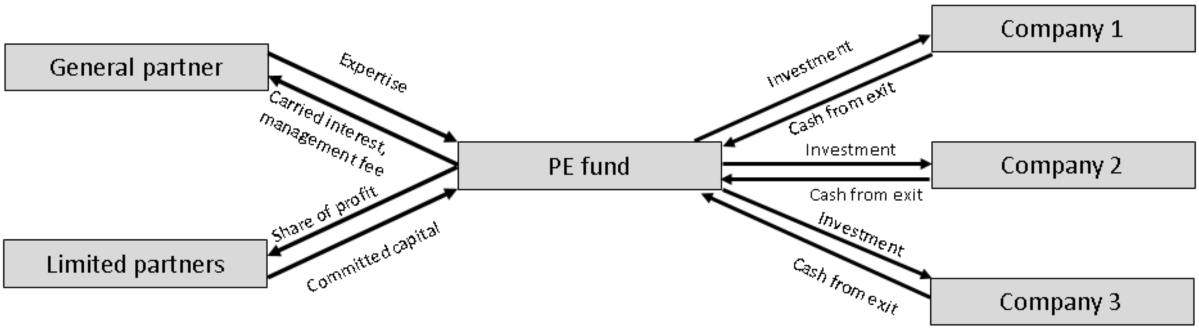


Figure 2: PE Fund Structure

3.1.2 Fund Lifecycle

Private equity funds' lifecycle involves four phases, the first of which is the fundraising phase which usually lasts in the interval of half a year to two years (Lerner, Leamon, & Hardymon, 2012). During this time period, the general partners look for limited partners who are interested in investing capital in the fund. Upon finding investors who are in total willing to invest an amount equal to the fund manager's target fund size, the fund is considered "closed", and new entrants are not allowed to directly invest in the fund. The existing investors are then referred to as primary limited partners, and the amount they have agreed to invest in the fund during its lifecycle is referred to as committed capital. Past this point, new investors can only enter the fund through purchasing a stake from the existing primary investors in what is considered a secondary transaction.

Beyond the fundraising phase, the investing phase starts, which can take up to six years (Lerner, Leamon, & Hardymon, 2012). In this phase, the general partner makes contribution calls which require the limited partners to send in a part of their committed capital to the fund, while the general partner finds and conducts private investments into companies. The companies can either be privately held to begin with, or publicly traded, in which case the private equity fund delists the company after purchasing it. Next, the holding phase begins, which involves the fund management using their expertise to actively manage their portfolio companies and direct them towards operational improvement, maximizing the value of the portfolio companies. The final phase is the divestment phase, where the general partner exits the portfolio companies, and distributes to the limited partners their share of the fund's capital. Most often used exit strategies in this phase involve an Initial Public Offering (IPO), selling the company to a strategic buyer, i.e., an industry competitor, or performing a secondary buyout, i.e., selling the company to a different private equity fund. During both the investment and divestment phase, the valuation of portfolio companies which are first bought and then exited is based on an EV/EBITDA valuation multiple at transaction date for Buyout funds. In contrast, as Venture funds invest in earlier stage companies which are not yet profitable, the valuation of their portfolio companies at transaction date is based on an EV/Revenue multiple. The obvious exception from this involves exiting portfolio companies with an IPO, where the companies are valued to their share price on the market for both Buyout and Venture funds. After limited partners have received their share of the fund's capital, the fund is considered liquidated. Figure 3 summarizes the contributions and distributions of capital to limited partners during a private equity fund's lifecycle, which due to its shape bears the name J-curve.

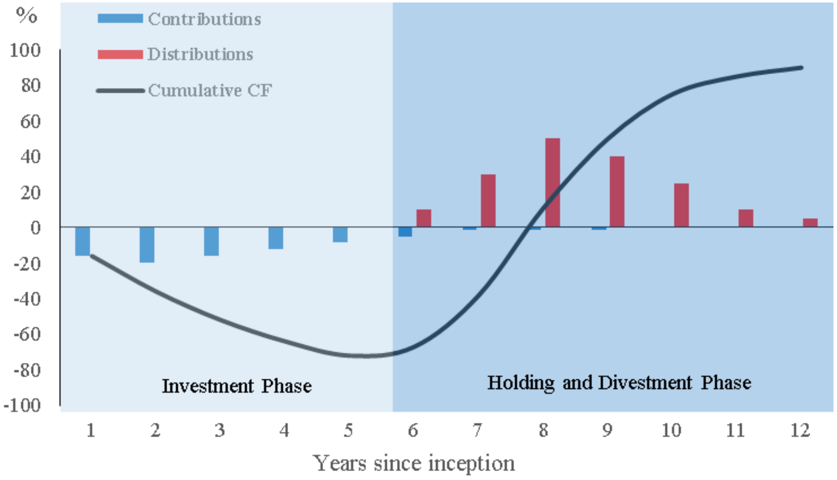


Figure 3: Fund Lifecycle's J-curve

Upon liquidating a successful fund, general partners often launch a new fund with the same investment theme, although this can happen before the previous fund's liquidation. This gives a sequence of fund series, which is a series of funds from the same fund manager with the same investment theme. These funds often include a sequence number in their name. Additionally, the fund manager can choose to expand into a different investment theme, giving rise to multiple fund series.

3.1.3 Measures of Performance

Net IRR (NIRR)

The NIRR is the most used measure of evaluating returns by limited partners and general partners (PitchBook, 2022a). IRR is defined as the discount rate which makes the net present value (NPV) of future cash flows to limited partners equal to zero. While Gross IRR is the IRR before deducting fund managements' fees and carried interest, Net IRR is the IRR when these components of management compensation have been subtracted from cash flows. As such, the calculation of NIRR can be presented through the following formula:

$$NPV = \sum_{n=0}^{N} \frac{Distributions_n - Contributions_n - ManagementFee_n - CarriedInterest_n}{(1 + NIRR)^n}$$

Where:

$$N = Length\ of\ the\ fund's\ lifecycle$$

Essentially, NIRR is similar to annualized return which is used for evaluating public markets returns, with the exception being that unlike annualized return, NIRR accounts for irregularities in terms of timing and size of cash flows (PitchBook, 2022a). Nevertheless, the measure has some drawdowns, the first of which being that it assumes the distributions to the investors are reinvested at the same rate of returns along the time horizon, which is not aways realistic. Additionally, NIRR does not show the total return of the investment. Thus, while a fund might display a high NIRR, if its lifetime only lasted for 3 years, the total change in the invested amount might not be too significant, which would not be captured through the NIRR measure.

TVPI

An approach for reducing the drawdowns of the NIRR measure is by supplementing it with the use of a Cash Multiple, which shows the total return of the investment (PitchBook, 2022b). The most commonly used Cash Multiples for evaluating private equity returns include DPI, RVPI and TVPI. The first from the list, Distributions to Paid-In (DPI), is calculated by dividing the distributed capital by the capital contributed by investors:

$$DPI = \frac{Total\ distributions}{Total\ contributions}$$

Further, Remaining Value to Paid-In (RVPI) shows the expected future distributions divided by the capital contributed by investors. Expected future distributions are assumed to be the unrealized value which remains in the fund, also referred to as Net Asset Value (NAV):

$$RVPI = \frac{Unrealized\ value\ of\ fund\ (NAV)}{Total\ contributions}$$

The Total Value to Paid-In (TVPI) is the sum of DPI and RVPI. Hence, it shows total value which investors have received and can expect to receive from the fund before the end of the fund's lifecycle compared to the invested amount. As a fund moves closer to the end of its lifecycle, it exits its unrealized investments, meaning the RVPI will decrease, transferring its value to DPI, which increases. TVPI and its components are summarized in the formula below:

$$TVPI = DPI + RVPI = \frac{Total\ distributions + Unrealized\ value\ of\ fund}{Total\ contributions}$$

PME

An alternative to traditional returns measures in private equity is the Public Market Equivalent (PME) (PitchBook, 2022a). PME comes in different variants, and was developed to allow for comparing private equity returns to that of public markets. Originally, due to private equity investments being illiquid and the timing of them being inconsistent, comparing them to public markets would become difficult. PME deals with this issue by handling limited partner contributions as purchases of shares in a stock market index, and distributions as sale of these shares. By doing so, PME shows the market-adjusted returns from investing in a specific private equity fund (Harris, Jenkinson, Kaplan, & Stucke, 2022). A commonly used approach for converting IRR to PME is through the PME+ variant developed by Rouvinez (2014). PME+ can be presented through the following formula:

$$PME_{PME+,T} = \sum_{s=0}^{t} (contribution_s - \lambda_T * distribution_s) * \frac{I_t}{I_s}$$

Where:

$$I_s = value\ of\ index\ at\ date\ s$$

$$\lambda_T = \frac{(S_C - NAV_{PE}, T)}{S_d}$$

$$S_C = \sum_{s=0}^{T} \left(contribution_s * \frac{I_T}{I_s} \right)$$

$$S_d = \sum_{s=0}^{T} (distribution_s * \frac{I_T}{I_s})$$

Lambda is chosen such that:

$$NAV_{PME+,T} = NAV_{PE,T}$$

IRR is then calculated as cashflows:

$$PME+_T = IRR(Contributions, \lambda_T, Distributions, NAV_{PE,T})$$

Additionally, PME can be used to convert a Cash Multiple, such as TVPI, into a public market equivalent. This is most commonly done through the use of Kaplan Schoar PME, developed for this purpose by Steve Kaplan and Antoinette Schoar (2005). This version of PME can be calculated through the following approach:

$$KS - PME = \frac{FV(Distributions)}{FV(Contributions)}$$

With:

$$FV(Distributions) = \sum_{t} (Distributions(t) * \frac{I_T}{I_t})$$

$$FV(Contributions) = \sum_{t} (Contributions(t) * \frac{I_T}{I_t})$$

$$I_s = value\ of\ index\ at\ date\ s$$

PME for its part also has numerous limitations. One of these involves that the measure only shows the value of a private equity investment if these funds were invested in a stock market index, however, without comparison to the actual IRR of the fund, this measure does not give extensive insights and cannot be used for a private equity fund evaluation. Furthermore, PME+ calculates the resulting PME by scaling the distributions to fit with an investment in a stock market index. However, this comes at the expense of resulting cash flows from the index being mismatched to that of the private equity fund.

3.2 Machine Learning

In this paper, the methods used for prediction range from very simple to very complex. The are two reasons for this. Firstly, we want to quantify whether there is a continual improvement in prediction accuracy with more complex models, and if so, the extent to which this improvement takes place. Secondly, there exists a trade-off between flexibility and interpretability of regression models. The more complex a model is, the more flexibility is available, but it is also less interpretable. Therefore, estimation of a simple linear regression model will serve as an explanation of how the predictors we use interact with our response.

Another important trade-off to note is bias versus variance. The expected test mean squared error can be decomposed into variance, bias and variance of the error term (James, Witten, Hastie, & Tibshirani, 2021):

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon)$$

$$Bias[\hat{f}(x_0)] = E[Bias(\hat{f}(x_0)] - f(x)$$

$$Var[\hat{f}(x_0)] = Var[\hat{f}(x_0) - \hat{f}(x_0)]^2$$

With:

$$y = Vector\ of\ observed\ values\ of\ the\ variable\ being\ predicted$$

$$\hat{f}(x) = Predicted\ values$$

The variance in this case is the amount by which the estimated f would change if it was estimated on another dataset and since variance and bias are non-negative, variance cannot be lower than the irreducible error. Bias is the error introduced by approximating real-life problems. The way to illustrate this is by considering the two extreme cases. For instance, drawing a curve that passes through every single training observation will result in very low bias, but the variance will be very high. In contrast, fitting a horizontal line to the training data results in very high bias, but very low variance.

The importance of this concept can be illustrated when considering the effect which reduced bias has on the expected test error. The charts below show how with increased model complexity, the squared bias gradually decreases and the variance increases. The effect of this on train error and expected test error is modelled in Figure 4, which is presented below.

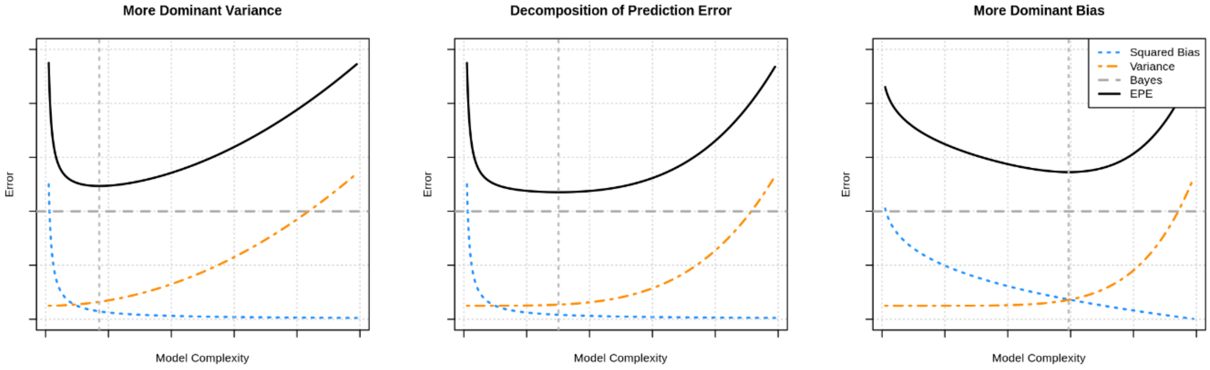


Figure 4a: Bias-Variance Trade-Off (Dalpiaz, 2018)

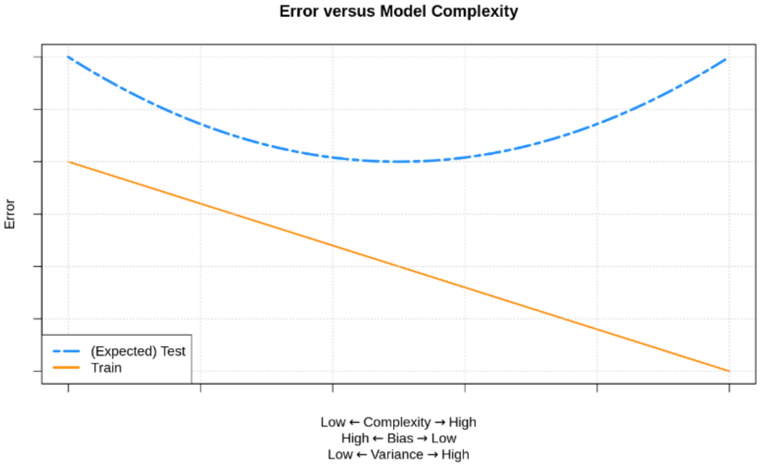


Figure 4b: Comparison of Train and Test Complexity (Dalpiaz, 2018)

Table 1 shows the result of estimating four different polynomial models with degrees 0, 1, 2 and 9. The mean squared error on the test set decreases with more polynomial degrees, until the model reaches 9 degrees, which yields a higher mean squared error than the model with only one polynomial degree. Table 1 also shows the effect that model complexity has on squared bias and variance, i.e., higher complexity gives a lower squared bias and higher variance. We expect to see similar results when we estimate models with increasing complexity and flexibility.

Degree	Squared Error	Squared Bias	Variance
0	0.0.22643	0.22476	0.00167
1	0.00829	0.00508	0.00322
2	0.00387	0.00005	0.00381
9	0.01019	0.00002	0.01017

Table 1: Bias-Variance Trade-Off Exemplified (Dalpiaz, 2018)

4. Data

Private equity data is generally quite difficult to obtain. Unlike for public funds, private equity funds are not required to report their data to the outside, other than to regulating bodies and to the funds' investors through quarterly reports and the PPM (Haves, 2022). Specialist data providers are able to gather private equity data through a combination of proprietary technology and direct requests for data submission from private equity fund managers and private investors (Brown, Harris, Tim, Kaplan, & Robinson, 2015). Private equity fund managers and investors then submit data to the specialist data providers on a voluntary basis. The challenge of obtaining such data for a researcher consists of the high costs charged by specialist providers, as well as uncertainty regarding the extensiveness and quality of data to be provided.

Based on an evaluation of the abovementioned factors among specialist data providers, we have chosen to use data from PitchBook – one of the largest data providers on private capital markets, as our primary source of data. Such selection was largely based on the breadth of variables available within the PitchBook platform, which will allow for a more in-depth analysis (PitchBook, 2022c).

4.1 Data Selection

While mark-to-market accounting standards require private equity fund managers to report the value of their funds at fair-value for the funds' investors, fund managers have significant freedom in choosing assumptions for calculating such fair-values (Easton, Larocque, & Stevens, 2021). This allows for fund managers to manipulate the reported returns for their funds to attract investors to new funds the managers plan to launch, or to hide significant decreases in fund value during economic downturns. When the funds are liquidated at the end of the funds' lifecycle, investors receive their final distributions of capital from the funds, meaning at this point there is no ambiguity about what final return the fund has generated. As such, we have selected to base our data sample exclusively on liquidated funds, to avoid the reported returns being manipulated.

Furthermore, we have chosen our data sample to include funds from fund types Buyout, Venture, and Growth Equity, as these are perceived as conventional private equity fund types (Lerner, Leamon, & Hardymon, 2012). Additionally, we have chosen to include funds from all vintage years available in order to maximize the statistical significance of our analysis.

4.1.1 Output Variable

Upon examining our data sample, which will be presented in full in the following sections, we observed that multiple funds lack data on one of NIRR or TVPI. Therefore, including both as output variables, or using both to calculate PME would significantly reduce our data sample, creating a probability that some potentially significant variables would become statistically insignificant. Based on this, we have chosen to use only one output variable for our analysis. For determining our metric of choice, we have compared the performance measures described in the Theory section of this paper. While NIRR and TVPI have certain limitations, such as NIRR having an underlying assumption that the distributions to investors are reinvested at the same rate of returns along the time horizon, which is not aways realistic, these measures continue being most frequently used returns measures by general- and limited partners (PitchBook, 2022b). Furthermore, due to only being provided raw data on NIRR and TVPI returns in our data sample, we lack information on the size and timing of contributions and distributions, which makes calculating PME values quite difficult. As such, PME variables will not be our metric of choice. Furthermore, while keeping in mind this paper's purpose of making the private equity market more accessible to retail investors, the NIRR measure is more easily comparable to annualized returns used for public markets than TVPI is. Therefore, we have selected to use NIRR as our output variable of choice.

As mentioned in the Theory section, a vocal limitation of NIRR as a measure of performance is its underlying assumption that distributions to the investors are reinvested at the same rate of returns along the time horizon, which is not always realistic. A potential option to address this limitation could involve adjusting NIRR such that distributions to the investors are reinvested at a different rate, for instance by reinvesting the distribution at the rate of the private equity firm's cost of capital. However, as mentioned in the paragraph above, our data sample lacks information on the size and timing of distributions, only providing raw data on the value of NIRR, making it unfeasible to adjust the rate of reinvesting distributions to investors. As such, we will use the raw NIRR figures for measuring fund performance.

Furthermore, as the available data sample contains funds from multiple vintage years as well as from multiple fund types, the resulting funds have quite different lifecycles and investment periods, meaning cash contributions and distributions for the funds would be quite different for each year, making the NIRR by year largely incomparable between funds. While an option could be to select all funds which have an equally long lifecycle, this would severely limit the size of our data sample. Consequently, we have chosen to base our returns data on NIRR received upon fund liquidation. This implies our analysis will forecast the total return fund investors can expect to receive from the moment of making a primary investment into the fund, until the end of the fund's lifecycle.

4.1.2 Predictor Variables

Our choice of predictor variables will largely be based on prior research of factors which influence performance of private equity funds, supplemented by own rationale for selecting proxies which represent these factors.

Fund Management's Experience:

Perhaps the most crucial factor impacting private equity fund performance involves the experience of the funds' management team. Kaplan and Schoar find that higher experienced fund managers are able to improve the funds' portfolio companies more efficiently, thus achieving higher returns for their funds (Kaplan & Schoar, 2005). To produce such analysis, the abovementioned researchers use the fund sequence as a proxy for management experience, concluding that a higher sequence number of funds has a positive effect on fund returns. We will therefore include this measure in our analysis through variable "Fund Series", which shows the fund's number in the sequence of funds from the same fund manager and same investment theme, as well as variable "Fund Number", showing the fund's number in the sequence of overall funds from the same fund manager. Furthermore, based on an analysis of North American private equity funds by Roggi and Giannozzi, the positive relationship between fund sequence and fund returns is only significant at high fund sequence numbers (Roggi & Giannozzi, 2019). To account for this convex effect of fund sequence on fund returns, we will therefore additionally use the logarithm of variables "Fund Series" and "Fund Number" as predictor variables.

Another highly significant predictor factor includes the extensiveness of the network of the fund's management team. Hochberg et al. examine the effect of private equity funds' networks on their performance, coming to the conclusion that funds with a stronger network have access to a larger number of potential deals, and are able to pool information from their network to produce better deal selection, thereby delivering higher returns for their funds (Hochberg, Ljungqvist, & Lu, 2007). Private equity firms account for this effect by having requirements of their employees having previously worked a few years within the investment banking industry, such that when they eventually climb the career ladder and become partners, they have a stronger network base they can use to originate deals for the funds they participate in (CFA Institute, 2022). As each new partner in a private equity fund is required to have developed a network prior to being accepted as an employee in the fund, the fund's total network grows with each new partner added. To account for this factor, we will therefore use the variable "Number of partners" as a proxy for fund management's network. We can also expect the relationship to be concave, as when the fund management already includes a high number of partners, the marginal effect of adding a single new partner would be smaller than when adding a single new partner to a fund which only had one previous partner. To account for this, we will also include logarithm of "Number of partners" as a variable in our analysis.

Funds' Strategy:

The funds' investment strategy is another categorizing factor which differentiates private equity funds between themselves and has a high potential for driving returns. A fund's investment strategy can be split into multiple factors, where the first involves the fund's geographic and industry preference for the companies the fund will invest in. Researchers of the private equity industry have developed two opposing views on whether a fund's specialization in a geographic region and industry has a positive effect on the fund's performance. Supporters of the portfolio theory argue that when diversification is reduced, the number of potentially profitable investments is reduced for the fund as well, leading to weaker performance (Manigart, et al., 2002). On the contrary, the resource based theory presents a view that a higher level of specialization allows the fund to build a competitive advantage through increased understanding of its focus area, enabling the fund to support its portfolio companies more efficiently and produce higher returns (Christensen, 2007). In a study from 2006, Lossen finds a positive effect on private equity fund returns from higher industry diversification, and no significant effect from geographic diversification (Lossen, 2006). We believe that while both factors could be significant, the impact of geographic diversification on fund returns could potentially be explained by other factors, such as the macroeconomic environment of- or the PE market conditions of the geographic region for investment. To examine this further, we include categorical variables "Preferred Industry" and "Preferred Geography" in our analysis, each categorizing the funds by their area of specialization. Additionally, to account for the possibility of some geographic regions having few observations in the data sample, we have added variable "Preferred Geography Fourgroup", where the funds' geographic preferences are divided into categories Europe, North America, Global, which is a mixed preference between Europe and North America, and finally Emerging Markets.

Another aspect of a private equity fund's strategy involves the fund type. As discussed in the Theory section of this paper, private equity includes three main types – Buyouts, where the fund focuses on majority stake investments in mature companies, Venture, where the fund focuses on minority stake investment in early-stage companies, and Growth Equity, where the fund buys minority stakes in mature companies (Lerner, Leamon, & Hardymon, 2012). As a result of these characteristics, the three fund types contain different risk profiles, consequently leading to varying returns. We will incorporate the effect of varying fund types on returns in our analysis by including a categorical variable "Fund Type", which can contain categories Buyout, Venture, and Growth Equity.

Perhaps a more efficient and detailed approach to capturing the effect of the difference in fund type can be achieved through splitting the "Fund Type" variable into its two components – preference for a majority or minority stake and the size of the companies which the fund will invest in. Loos (2006) explored the effect of ownership stake on private equity funds' returns, coming to conclusion that funds which prefer majority stakes for their investments are able to achieve superior performance. The study suggests that this could be a result of majority stakes allowing more experienced fund managers to apply their experience more effectively in the management of their portfolio companies, hence leading to stronger strategic improvement of their companies and higher returns. We incorporate this effect through a categorical variable "Majority/Minority Position", which can be equal to preferred majority stake, preferred minority stake, and no preference. Furthermore, Loos observes a convex relationship between the size of the companies which private equity funds invest in, and returns achieved by these funds, with investments in small- and mega-cap categories displaying increased returns for the funds compared to investments in mid- and large-cap categories. Potential explanations for this effect could include mega deals being generally undertaken by the largest and most successful funds, while companies from the small-cap category are at the start of their lifecycle and hence have the highest potential for growth, enabling successful fund managers to guide them on such growth trajectory, achieving higher returns. We include this convex effect by incorporating the variable "Average Deal Size" and the logarithm of said variable.

A further categorization of the funds' strategy involves the fund size. Roggi presents findings suggesting a concave relationship between fund size and fund returns (Roggi & Giannozzi, 2019). A potential explanation for such effect entails that past a given level, increasing the fund size produces dis-economies of scale, where it becomes difficult for fund managers to find enough strong investment opportunities to invest the entirety of the funds capital. Our variables to cover this concave effect will be "Fund Size", and the logarithm of "Fund Size".

The final factor categorizing a fund's strategy which we will examine includes whether a fund is ESG-focused. Unlike the previous factors, this is a factor which has not received sufficient research, mainly due to ESG-focused funds only recently beginning to increase in popularity. Such funds are characterized by them making impact investments which must contribute to achieving ESG goals (Esty & Cort, 2020). While such strategy might attract a larger amount of investors, particularly with investors' increasing demand for ESG positive investments, a focus on impact investments irrespective of whether such investments maximize financial return would lead to decreased fund performance. We therefore expect ESG-focused funds to underperform funds without a declared ESG-focus. To incorporate this factor, we use the categorical variable "ESG Focus", which takes values ESG focus and no ESG focus.

PE Market Variables:

The overall conditions of the private equity market have a strong impact on the performance of funds operating there, and must hence be accounted for. Private equity funds face different risks at each phase of a fund's lifecycle. As discussed in the Theory section, private equity funds need funding at the beginning of their lifecycle in order to close the funds and have cash to invest. Inability to close the fund at the target size might lead to fund managers not having sufficient funds for the investment phase and therefore being forced to miss out on profitable deals. Hence, a successful fundraising phase is an essential value driver. In a study from 2014, Duong confirms this effect by presenting empirical evidence backing a positive relationship between successful fundraising and private equity returns (Duong, 2014). We include this effect through variable "Funds Raised Market" which shows total number of funds raised in the market by year. The rationale behind this involves a smaller number of funds raised being an indicator of the difficulty of raising a fund during that particular year. We can expect that the relationship between fundraising and fund returns is convex, as the total fundraising being increased by one at already very high values cannot be expected to improve a specific fund managers chances of closing a fund during that year as much as total fundraising being increased by one at lower total fundraising values. Therefore, we will include a possible convex effect by adding the logarithm of “Funds Raised Market” to our analysis.

Further in the funds’ lifecycle, funds need to find high potential deals to invest in and later find buyers to exit said investments. For this, a favourable market environment for deals is crucial. Ljungqvist analysed the effect of deal market conditions on private equity fund returns, with results suggesting larger M&A volume leads to less fierce competition, allowing for fund managers to more easily find promising deals, in turn leading to higher private equity fund returns (Ljungqvist & Richardson, 2003). Additionally, Achleitner conducted a study of value creation drivers in private equity funds, coming to the conclusion that a significant portion of private equity fund value creation comes as a result of expansion in valuation multiples (Achleitner, Braun, Engel, Figge, & Tappeiner, 2010). To incorporate these effects, we include variables “M&A Volume (Capital Invested, MUSD)” and “M&A Volume (Deal Count)” which show total M&A volume in the market for each year, as well as variables “EV/Revenue” and “EV/EBITDA”, which show the said valuation multiples as market average per year. Additionally, we will incorporate the variable “Dry Powder Market”, which shows, on a market level, the total cash private equity funds can still request from the limited partners, as similarly to M&A volume, funds having large amounts of cash saved can be another measure facilitating for a favourable dealmaking market. For a similar reason to that for fundraising, we expect M&A Volume and Dry Powder to have a convex relationship with fund returns, as we expect the marginal effect of increasing M&A Volume or Dry Powder at already very high values to improve dealmaking opportunities for fund managers less than marginally increasing M&A Volume or Dry Powder at less high values. We will hence include the logarithm of variables “M&A Volume (Capital Invested, MUSD)”, “M&A Volume (Deal Count)” and “Dry Powder Market” in our analysis to incorporate this potential convex effect.

We take into account the geographical focus of the funds by obtaining geographically focused data for each of the mentioned PE market variables. The geographically focused data obtained consists of three categories – US based data, which will be applied to North American focused funds, Europe based data, which will be applied to European based funds, Global based data, which is a simple average of the North American and European based data and which will be applied to funds with a mixed preference for Europe and North America, and finally Emerging Markets based data, which will be applied to remaining funds.

Macroeconomic Variables:

Private equity performance is highly pro-cyclical, meaning the returns delivered by private equity funds throughout a year are positively correlated with the macroeconomic conditions during the same year (Phalippou & Zollo, 2005). Despite this, Aigner finds that the level of GDP growth and global stock market returns at the funds' inception have a negative impact on the returns the funds deliver by the end of their lifecycle (Aigner, et al., 2008). A possible rationale behind this is that positive macroeconomic conditions at the funds' inception drive up valuation when the funds make their investment, resulting in funds exiting their investments when the economic cycle has progressed into a downturn with lower valuations, hence leading to funds delivering a weaker return. The researcher also finds that the level of interest rates at the funds' inception negatively affect the funds' return delivered by the end of the funds' lifecycle. This is highly rational, especially for Buyout funds which tend to fund a significant amount of their investments with leverage. Furthermore, Phalippou observes a similar negative relationship between fund performance and the level of credit spread at the time private equity funds make their investments. For said study, the researcher used the difference between the yield of 10-year trusury bonds and BAA-rated bonds as a measure of credit spread. This result makes logical sense, as such credit spread can serve as a proxy for the level of risk in the economy, and with a higher level of risk one can intuitively expect weaker returns.

We incorporate the mentioned macroeconomic effects in our analaysis throgh variables "Stock Market Returns", "Bond Yield", "Credit Spread", and "GDP Growth", with all variables being based on yearly values. Simiarly to PE market variables, we take into account the geographical focus of the funds by obtaining geographically focused data for each of the mentioned macroeconomic variables. To do so, we select our proxy for "Stock Market Returns" to be the S&P500 for North America, Stoxx600 for Europe, and MSCI Emerging Markets Index for Emerging Markets. For the "Bond Yield" variable we have selected proxies of the US 10-year trasury yield for North America, the ECB's Euro area 10-year governemnt bond yield for Europe, and Bloomberg's Emerging Markets Bond Index for Emerging Markets. Furthermore, our selected proxies for "Credit Spread" include the ICE BofA's US High Yield Spread Index for North America, the ICE BofA's Euro High Yield Spread Index for Europe, and the ICE BofA's High Yield Emerging Markets Corporate Plus Spread Index for Emerging Markets. Finally, we have selected proxies for variable "GDP Growth" by collecting data from IMF on a US basis for North America, on a European basis for Europe, and on an Emerging

Markets basis for Emerging Markets. We have based our choice of proxies for macroeconomic variables on how many years back the proxies' macroeconomic data goes, which has been done with the purpose of keeping a largest possible number of observations in our dataset when connecting the macroeconomic data to fund returns.

4.2 Data Transformation

Upon completing the selection of our output and predictor variables, the subsequent step involves cleaning and transforming the data such that it is ready for the analysis. This includes removing missing observations, standardizing categorical variables, scaling numerical variables, and connecting the PE market- and macroeconomic variables to the output variable. While starting from a raw dataset of 1275 funds, the sample size was reduced to 600 funds. The general steps conducted for this include: 1) removing all funds with missing observations for output and predictor variables, 2) removing all funds where calculating the logarithmic variables produced an error due to the base variable for calculating the logarithm being equal to zero , 3) and removing funds which were of different fund type than Buyout, Venture, or Growth Equity. The following sections describe the more specific steps done for data transformation inside certain variables.

4.2.1 Output Variable

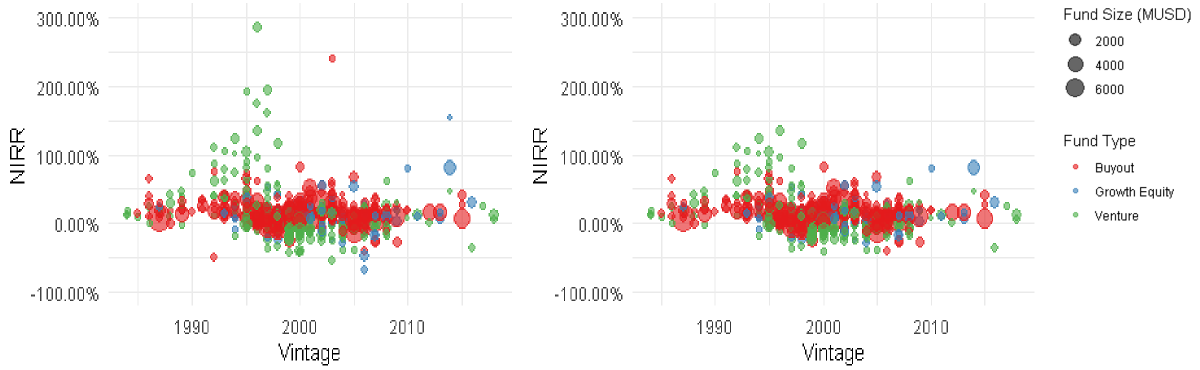


Figure 5: NIRR distribution before and after trimming outliers

The left side of Figure 5 presents an overview of NIRR from our data sample plotted against vintage years. Upon examining the distribution of NIRR, it is clear that it includes a few outliers, with a small number of funds achieving returns in the range of 200%-300%. Such returns can be driven by a single one of the fund's investment performing exceptionally well and bringing up the value of the entire fund, which occasionally happens for private equity funds (Lerner, Leamon, & Hardymon, 2012). As the occurrence of the abovementioned event is quite rare, it is not representative for the other funds in the overall data sample. Consecutively, predicting these events is highly difficult, and including outlier funds with returns driven by a single investment performing exceptionally well into our models would significantly decrease the models' predictive power. We have therefore addressed this issue by trimming NIRR at a 99% confidence interval. The right side of Figure 5 presents an illustration of the data sample post-trimming of NIRR, and as seen from the figure, this yields a much more even distribution of NIRR.

4.2.2. Categorical Variables

The raw data for categorical variables "Preferred Industry" and "Preferred Geography" were initially formatted in a way which made it difficult to include this data in the analysis. This involved each observation of the two variables consisting of a list of countries or industries where each fund prefers to conduct its investments. For instance, while one fund could have "Preferred Geography" equal to "Norway, Sweden", another could have this variable equal to "Norway, Denmark". While there are clear similarities in the values of the variable for the two funds, an algorithmic analysis would see it as two distinctively different observations. The same applies to values of the variable "Preferred Industry". Our solution is to group observation of each variable into more generic categories, such that the variables are no longer formatted as lists. For the "Preferred Industry" variable we have used GICS industry taxonomy to do so, as it is among the most widely accepted approaches to industry classification by finance professionals (MSCI, 2022). In applying the GICS industry taxonomy to the variable "Preferred Industry", we have grouped each observation of the variable, with the list of values it previously contained, into industry groups defined in the GICS.

For the "Preferred Geography" variable, we have used the UN geoscheme for classification of global geographic regions, due to this being a highly used approach for region classification in statistical analysis (United Nations, 2022). Upon examining the data sample further, we have observed that a large number of funds have geographic preferences for subregions of Europe and the US, and as such we have incorporated this effect through a further breakdown of US and Europe. For Europe, we have again used the UN geoscheme for Europe, except with applying minor changes (United Nations, 2022). These changes include classifying the Baltic countries as part of Eastern Europe, due to these countries historically having closer economic ties with Eastern Europe than the Nordics. A further change involves including the

UK as part of Western Europe instead of the Nordics, due to the country being both geographically and economically more connected to the former. As a final change, Greenland has been classified as part of Nordics instead of North America. This was based on a detailed examination of the funds with a geographic focus on North America, which led us to a conclusion that these funds generally prefer investments in the US, and due to strong economic differences between US and Greenland, inclusion of the latter in the group of North America could reduce the statistical significance of this categorical variable. For the US, we have used the US Census Bureau's classification of US regions (United States Census Bureau, 2022a). This is due to the mentioned classification being among the most widely accepted approaches to geographic classification by data collectors (United States Census Bureau, 2022b).

4.2.3. PE Market Variables and Macroeconomic Variables

We previously chose to select our output variable to be funds' NIRR at liquidation instead of NIRR per year, while PE market variables and macroeconomic variables were collected on a yearly basis. As such, we need to connect the latter to the output variable by assigning each fund a single observation of PE market variables and macroeconomic variables. While we have data on PE market variables and macroeconomic variables for any year of each fund's lifecycle, it is highly unrealistic to expect an investor to forecast these values with high precision at the beginning of the fund's life. Therefore, it intuitively makes sense to assign each fund a single observation of historical PE market variables and macroeconomic variables, i.e., a value of these variables before the fund's inception.

In order to decide from which year relative to the funds' inception the PE market variables will be selected, we have run a regression of PE market variables against the NIRR, as results of said regression will indicate how well PE market variables from given year relative to fund inception suit to predict private equity fund performance. For the regression we have selected year relative to fund inception ranging from 4 year before funds' inception, to the year of the funds' inception. As an example, "Funds raised 1" involves the value of "Funds Raised Market" during the year of the funds' inception, while "Funds raised 2" involves the value of same variable one year before the funds' inception. Appendix B.1 illustrates the results of the regression. From the regression it appears that none of the PE market variables are statistically significant for NIRR at liquidation. However, it must be kept in mind that these variables seem statistically insignificant in the current combination of them in the specific regression model, which does not mean these variables will necessarily be insignificant when included in the final model in our analysis. As we can also see, with an MSE of 0.083, PE market variables have significant predictor power towards explaining the variations in NIRR, especially when taking into account the MSE of our final model. Without having statistical evidence on which year relative to funds' inception is most optimal to select for PE market variables, we have selected the year of the funds' inception, as it intuitively makes sense that most recent data is a stronger predictor than data from any year prior to that.

We ran a similar regression for macroeconomic variables against NIRR, with the results of the regression presented in Appendix B.1. In line with the results from the PE market variables regression, the macroeconomic variables regression delivers a solid MSE of 0.086. Further, it is clear that for all macroeconomic variables except "Bond Yield", the year of the funds' inception is the only statistically significant year. For "Bond Yield", the year prior to the year of the funds' inception is the sole statistically significant year. Therefore we select the year prior to the year of the funds' inception for "Bond Yield", while selecting the year of the funds' inception for the remainder of the macroeconomic variables. The PE market data and the macroeconomic data is then connected to the output variable.

4.3 Dataset Overview

Our final dataset includes 600 private equity funds, which are split between 314 Buyout funds, 220 Venture funds, and 66 Growth Equity funds. The funds have vintage years in the interval of 1984 to 2018, with the majority centred around year 2000, resulting in a fairly normal distribution of vintage years, as seen in Appendix A.1. The data sample predominantly consists of funds focusing on investments in North America, with 321 funds belonging to this category, as well as 74 funds focusing on investments in Europe, 178 funds preferring a combination of Europe and North America, and 27 focusing on the rest of the world.

Table 2a illustrates a statistical summary of the numerical variables in our dataset. Some important features that need mentioning involve the difference between mean and median for the majority of numerical variables. Upon examining the distribution of numerical variables in Appendix A.1, it is clear that a majority of such variables have a somewhat skewed distribution towards larger values. For "Fund Management's Experience" and "Fund Strategy " this is likely due to variables not being able to inherit negative values, while positive values larger than the mean can occur. Meanwhile, when examining the historical development of PE market variables, we can see that such variables increase in value exponentially over the historical period, which provides the most reasonable explanation for the skewed distribution towards larger values. The significant tail towards larger values results in a high standard deviation for these numerical variables. It is worth noting that transforming variables into the logarithm of corresponding variables yields a more even distribution, hence resolving the potential issue of initially having a distribution with partial departures from normal distribution.

Variables	Mean	Median	Standard dev.	Data source
Output variable				
NIRR	0.14	0.11	0.23	PitchBook
Fund Management's Experience				
Fund Number	4	3	5	PitchBook
Fund Series	3	3	2	PitchBook
Number of Partners	8	6	6	PitchBook
log_fundnumber	1.06	1.10	0.84	PitchBook
log_fundseries	0.78	0.69	0.69	PitchBook
log_numberofpartners	1.76	1.75	0.76	PitchBook
Funds' strategy				PitchBook
Fund Size (MUSD)	471	216	746	PitchBook
Average Deal Size (MUSD)	42	14	131	PitchBook
log_fundsize	5.34	5.37	1.34	PitchBook
log_averagedealsize	2.66	2.60	1.36	PitchBook
PE market variables (yearly)				PitchBook
Funds Raised Market	900	481	1,110	PitchBook
Dry Powder Market (MUSD)	21,277	3,835	44,632	PitchBook
M&A Volume (Cap. invested, MUSD)	1,143,936	30,329	1,956,856	PitchBook
M&A Volume (Deal count)	22,981	960	43,945	PitchBook
EV/Revenue	1.06	1.09	0.38	PitchBook
EV/EBITDA	9.46	9.93	2.41	PitchBook
log_fundsraisedmarket	5.53	5.37	1.34	PitchBook
log_drypowdermarket	7.20	7.58	1.51	PitchBook
log_M&Acapitalinvested	12.59	12.87	1.39	PitchBook
log_M&Adealcount	8.42	8.61	1.09	PitchBook
Macroeconomic variables (yearly)				PitchBook
Stock Market Returns	0.09	0.09	0.24	Investing.com, MSCI [1]
Bond Yield	0.04	0.03	0.05	Investing.com, ECB, Bloomberg [2]
Credit Spread	0.05	0.04	0.04	FRED St. Louis [3]
GDP Growth	0.03	0.04	0.02	IMF [4]

Table 2a: Dataset Overview – Numerical Variables

Variables	Categories	Description	Data source
Funds' strategy			
Fund Type	3	Includes fund type categories: Buyout, Venture, and Growth Equity	PitchBook
Preferred Industry	19	Includes industry groups: Banks, Chemicals, Commercial and Professional Services (B2B), Consumer Durables & Apparel, Consumer Non-Durables, Consumer Services (B2C), Energy, Financial Services, Insurance, Materials, Media & Entertainment, Pharmaceuticals Biotechnology & Life Sciences, Retailing, Software & Services, Technology Hardware & Equipment, Telecommunication Services, Transportation, and No preference.	PitchBook
Preferred Geography	17	Includes geographical regions: United States, Midwest United States, Northeast United States, Southeast United States, Westcoast United States, Canada, North America, Pan-European, Eastern Europe/Russia, Nordics, Southern Europe, Western Europe, Africa, Asia, Latin America, Oceania, and No preference.	PitchBook
Preferred Geography Fourgroup	4	Pooled version of the Preferred Geography variable. Pan-European, Eastern-European/Russian, Northern, Southern European, and Western European funds are categorized as "Europe". Funds in United States and Canada are categorized as "North America". Funds aiming to invest in a combnation of Europe and North America are categorized as "Global". Remaining funds are categorized as "Emerging Markets".	PitchBook
Majority/Minority Position	3	Includes fund managers' preference for getting a controlling position in the deals they will conduct. The categories are: preference for a majority position, preference for a minority position, and no preference.	PitchBook
ESG Focus	2	Is equal to "ESG focus" if the fund has stated they will focus on ESG impact investments through their deals, and is equal to "No ESG focus" otherwise.	PitchBook

Table 2b: Dataset Overview – Categorical Variables

Table 2b adds an overview of our used categorical variables. An important point to notice is again the distribution of the variables. This is especially vital for variable "ESG Focus", for which, as illustrated in Appendix A.1, only 1% of the funds in the data sample are categorized as ESG-focused, while the remainder are not. While less extreme, such uneven distribution is also seen for some other categories including certain industries in the "Preferred Industry" variable and certain geographic regions in the "Preferred Geography" variable having only a few observations, as seen in Appendix A.1. Such uneven distribution can result in the corresponding categories with a smaller number of observations failing to become significant in the analysis.

5. Methodology

5.1 Test-Train and Overfitting

To validate the robustness of the models we are going to estimate, we will divide the data we have into a train set and a test set on an 80-20% basis. The reason for this split is to approximate unobserved real-world values as accurately as possible. It is not possible to conclude on the accuracy of our models based on their performance for the observed data, which is why a random subset of the original dataset is drawn out before we build our models. Thus, we can use the models to predict on a dataset that the model has not trained on and which is therefore “unobserved”.

The out-of-sample error is also called a generalization error (James, Witten, Hastie, & Tibshirani, 2021). This is due to this error being caused by model failing to generalize patterns seen on the training data to future data. There are three key ideas concerning the train-test split. The out-of-sample error is what one puts most emphasis on, as this is the best approximation of the prediction accuracy on new data. The in-sample-error is almost always smaller than out-of-sample error. Finally, the reason for the former having a smaller error is due to overfitting.

Overfitting is a behavior in machine learning which occurs when the machine learning model can predict well on the training set, but not on new data. This happens because the model is not able to generalize the predictions of data and fits the estimate too closely to the training set, producing inaccurate predictions on the test set, as illustrated in Figure 6. There are three categorizes of causes for overfitting. Firstly, and most important for this paper, is noise learning on the training set. This happens because the dataset on which the model is trained is too small in size, which makes it difficult for the model to separate representative data from noise.

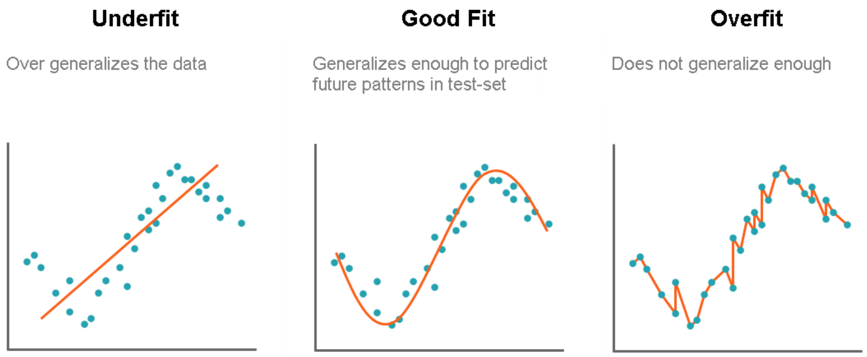


Figure 6: Machine Learning Algorithm Overfitting Data (Corporate Finance Institute, 2022)

Secondly, the concept of hypothesis complexity can also explain why some models are overfit. There is a fundamental trade-off in statistics and machine learning between variance and bias. As explained in the Theory section, when increasing complexity, a decrease in squared bias is expected, but at the cost of increased variance. By including more variables to the model, the algorithms will create too many sets of hypotheses and increase its complexity. This will subsequently lead to more variance, i.e., more accuracy on training data, but lower accuracy on out-of-sample data. This is our main motivation for using stepwise selection, which we will introduce in the following subsections, as it reduces number of variables, and hence, by using said method, we can expect higher accuracy and increased interpretability.

Lastly, an often used method for selecting the best model involves multiple comparisons procedures that are ubiquitous in both induction algorithms and other Artificial Intelligence algorithms. In these processes, multiple models are compared based on scores from an evaluation function and one with the highest score is selected as the optimal model. However, the choosing process may in itself produce biases towards a type of model that is not the most accurate on out-of-sample data or find local minima leading to different solutions at each iteration. The next sections will introduce and explain our choices of models.

5.2 Ordinary Least Squares

Ordinary least squares is a method of estimating the parameters in a regression model by minimizing the sum of the squared residuals (James, Witten, Hastie, & Tibshirani, 2021). OLS uses a method of estimating a linear line through data points which results in the lowest sum of squared differences between the actual values in the dataset and the fitted values from the final model. The method involves that for a dataset:

$$x_i, y_{i=1} \subset \mathrm{X} \times \mathbb{R}$$

Where:

$$x_i = \ a\ vector\ of\ predictors$$

$$y_i = \ the\ response, or\ in\ our\ case: Net\ Internal\ Rate\ of\ Return$$

OLS models the response variable as a linear function of the predictors through the following function:

$$y_i = x_i \beta + \varepsilon_i$$

Where:

$$\beta = coefficients\ for\ predictor\ variables$$

$$\alpha = intercept$$

$$\varepsilon_i = \ unobserved\ scalar\ random\ variables\ which\ account\ for\ the\ difference\ between\ observed\ and\ predicted\ values$$

The OLS method, given that assumptions of multicollinearity, homoscedasticity and uncorrelated variance are fulfilled, provides the estimate of least variance and mean-unbiasedness when errors have finite variances. This makes it an ideal method for understanding the relationship between NIRR and the variables in our dataset, as it is the method which provides the most interpretability.

By looking at the sign of the coefficient of each parameter, we can see whether it has a positive or negative effect on the output variable, and by looking at the value of the coefficient, we can see the magnitude of this effect. Furthermore, the P-value of each coefficient tells us whether this relationship is significant. These P-values come from the null hypothesis that there is no correlation between the predictor and the response variable. Therefore, if the P-value is less than a chosen significance level, which in statistical analysis is often set to a 5% level, the sample data provides enough evidence to reject the null hypothesis for the entire population.

5.3 Stepwise Regression

To achieve a model which includes only the variables found to have a significant relationship with the predictor, stepwise regression is used. This is a construction of a regression model which iteratively selects independent variables to be used in the final optimal model (James, Witten, Hastie, & Tibshirani, 2021). At each iteration, the algorithm adds or removes a predictor and tests for significance. There are three main methods of stepwise regression. Firstly, forward selection begins with no variables and iteratively tests each variable for significance until an optimal solution is found. Backward elimination, however, starts at the other end with all variables and tests, by iteratively removing each variable to test if the removed variable is significant. Finally, best subset selection tests all possible combinations of variables in the model, however, this is computationally expensive and is effective for datasets with a smaller number of independent variables.

Weaknesses of forward selection and backward elimination include that they may produce different results depending on where they start, i.e., they may find local minima that are less optimal than the global solution. Best subset selection avoids this problem by testing all possible iterations of the model. This is, however, computationally expensive. Therefore, a hybrid model has been introduced as a compromise between these methods. It starts as a forward selection algorithm, except after adding each new variable, it may also remove any variable which no longer provides any additional improvement to the model. As such, we have chosen this method among our methods for estimating prediction models.

5.3.1 Selection Criteria

Each method described above needs a criterion for evaluating the quality of each model fit and to decide whether adding or removing a variable leads to an improvement. To evaluate whether one model performs better than another, its mean squared error (MSE) may be used to quantify the extent to which the predicted response value for all observations are close to the true values (James, Witten, Hastie, & Tibshirani, 2021). Mean squared error can by defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

Where:

$$\hat{f}(x_i) = \ prediction\ for\ the\ i - th\ observation\ provided\ by\ \hat{f}$$

$$n = \ total\ number\ of\ observations$$

As such, MSE shows the squared average difference between an actual observation and the prediction made by the model using independent variable values for that observation. MSE values generally appear in the range of 0 to ∞, with 0 implying that the model perfectly predicts the actual observations. However, an important facet of a related measure that is important to keep in mind is that RSS will always decrease as more variables are added to the model on the training data. Because MSE = RSS/n, this means that MSE cannot be used as a measure of quality of fit. A solution for this is to adjust the training error by the increase in model size. One of the ways one can adjust the error by size is the Akaike Information Criterion.

5.3.2 Akaike Information Criterion

The Akaike Information Criterion is an estimator of prediction error and thereby of the relative quality of the statistical models for a given set of data (James, Witten, Hastie, & Tibshirani, 2021). Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models, aiming to maximize this value. AIC is defined as:

$$AIC = 1/n * (RSS + 2d\hat{\sigma}^2)$$

The foundation for the creation of the criterion is information theory. Under this, a statistical model is always expected to lose some information inherent in the dataset when it is used as a representation of the processes that generated that data. The criterion calculates an estimate of the relative amount of information lost by each model, and the less information lost, the higher quality of the model. It does so by rewarding the goodness of fit (RSS) but also penalizing each model by a function of the number of estimated parameters. Thus, it encourages goodness of fit but also discourages overfitting. An important quality to keep in mind when using this criterion is that it provides no measure of absolute quality, only the quality of a model relative to other models. Therefore, all models may have a very low quality of fit without warning.

5.4 Ridge Regression

Ridge regression is another way of performing variable selection by “shrinking” variables according to a chosen parameter towards zero based on their explanatory power as shown in the equation below (James, Witten, Hastie, & Tibshirani, 2021). This may produce lower mean squared error and increased explanatory power as compared to OLS. The Ridge method can be described through the following equation:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

Where:

$$\lambda = penalty\,parameter$$

Larger values of lambda lead to more variables near zero. We include variable selection in this regression to further simplify the model. To validate the accuracy of the models, leave-one-out cross-validation (LOOCV) is used to estimate the sum of mean squared errors for each model. The LOOCV method can be presented as follows:

For a dataset:

$$x_i, y_{i=1}^{n} \subset X \times \mathbb{R}$$

Estimate a function:

$$\hat{f}(x) = \alpha^{\intercal}\phi(x)$$

Such that

$$\frac{1}{n}\sum_{i=1}^{n}(\alpha^{\intercal}\phi(x_i) - y_i)^2 + \lambda\alpha^{\intercal}\alpha$$

Is minimized. The error is then computed as:

$$E_{LOOCV} = \frac{1}{n}\left|\left|H\widetilde{H}^{-1}y\right|\right|^2$$

$$H = I - \Phi^{\intercal}L$$

$$\Phi = (\Phi(x_1)| \dots |\Phi(x_n))$$

$$L = (\Phi\Phi^{\intercal} + \lambda \mathrm{I})^{-1}\Phi$$

To estimate different models with increasing penalization, a grid of values is created equal to

$$K \coloneqq 10^{\{1,-10\}}$$

With leaps of about 0.1.

In contrast to ordinary least squares where only one set of coefficients is estimated, ridge regression estimates a different set of coefficient for each value of λ. It is therefore important to select the best value for λ. This is why we have chosen to use leave-one-out cross-validation in order to select the most ideal value for the tuning parameter. Although it is computationally expensive, using LOOCV to calculate the total cross-validation error is ideal results-wise, as it uses each observation as a test set, while the model is trained on the remaining observations. Thus, for a dataset of size n, n models are estimated to find optimal value of λ.

5.5 Support Vector Regression

Support vector regression (SVR) is the next method we use to estimate the performance of private equity funds. This model is significantly more complex and hence flexible. It is a constructive learning procedure rooted in statistical learning theory based on the principle of structural risk minimization, implying that it aims to minimize the bound on the generalization error, i.e., the error made by the learning machine on out-of-sample data, rather than minimizing the empirical error. This gives good generalization capability and means it usually perform well when applied to data outside the training set.

The main advantage of using support vector regression, compared to simpler methods, is that SVR is flexible through allowing the user to define the magnitude of error that is acceptable. Next, it estimates either a line or a hyperplane in higher dimension to fit the data. The main difference to OLS is that in SVR, it is not the error that is minimized in the objective function, but rather the coefficients:

$$MIN\frac{1}{2}||w||^2$$

Given:

$$|y_i - w_i x_i| \leq \varepsilon$$

Where:

$$w_i = weight\ for\ variables$$

Meaning that the within the constraint that the difference between the estimated value and the true value is less than a pre-defined maximum error, the coefficients are minimized. In our case we tune the value of epsilon by 100 different values between 0 and 1. Despite finding an optimal value of epsilon, there will still be errors which are larger than epsilon. These deviations, where the error is larger than the defined maximum error, can be summed and denoted as E. Thus, we can also control the total deviation by adding a term to the objective function:

$$MIN\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}|\mathrm{E}_i|$$

Where:

$$C = \ a\ hyperparameter\ which\ is\ also\ tuned\ before\ the\ final\ fit$$

In our case, the model is retrained for all values of C between 1 and 10. In order to find a hyperplane which fits the data in a higher dimensional space, we have used a radial basis function kernel.

5.5.1 Radial Basis Function Kernel

The radial basis function (RBF) is a popular kernel function used in various kernelized learning algorithms. It is the most generalized form of kernelization and most popular in use as it very similar to the Gaussian distribution. The RBF kernel on two samples $x \in \mathbb{R}^k$ and x', represented as feature vectors in some input space, is defined as

$$K(x, x') = exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

$||x - x'||^2$ may be recognized as the squared Euclidean distance between the two feature vectors. σ is a free parameter and represents the variance. An equivalent definition involves a parameter $\gamma = \frac{1}{2\sigma^2}$:

$$K(x, x') = exp\left(-\gamma||x - x'||^2\right)$$

Since the value of the RBF kernel decreases with distance and ranges between zero (in the limit) and one (when $x = x'$), it has a ready interpretation as a similarity measure. The feature space of the kernel has an infinite number of dimensions; for $\sigma = 1$, its expansion using the multinomial theorem is:

$$exp\left(-\frac{1}{2}||x - x'||^2\right) = exp\left(\frac{2}{2}x^T x' - \frac{1}{2}||x||^2 - \frac{1}{2}||x'||^2\right)$$

$$= \exp(x^T x')\exp\left(-\frac{1}{2}||x||^2\right)\exp\left(-\frac{1}{2}||x'||^2\right)$$

$$= \sum_{j=0}^{\infty}\frac{(x^T x')^j}{j!}\exp\left(-\frac{1}{2}||x||^2\right)\exp\left(-\frac{1}{2}||x||^2\right)$$

$$\sum_{j=0}^{\infty}\sum_{n_1+n_2+\cdots+n_k=j}\frac{exp\left(-\frac{1}{2}||x||^2\right)\left(x_1^{n_1}\dots x_k^{n_k}\right)}{\sqrt{n_1!\dots n_k!}}exp\left(-\frac{1}{2}||x'||^2\right)\left(\frac{x_1'^{n_1}\dots x_k'^{n_k}}{\sqrt{n_1!\dots n_k!}}\right)$$

$$= \langle\varphi(x), \varphi(x')\rangle$$

$$\varphi(x) = exp\left(-\frac{1}{2}||x||^2\right)\left(a_{l_0}^{(0)}, a_1^{(1)}, \dots, a_{l_1}^{(1)}, \dots, a_1^{(j)}, \dots, a_{l_j}^{(j)}, \dots\right)$$

Where $l_j = \binom{k + j - 1}{j}$

$$a_l^{(j)} = \frac{x_1^{n_1} \dots x_k^{n_k}}{\sqrt{(n_1! \dots n_k!)}} \mid n_1 + n_2 + \cdots + n_k = j \wedge 1 \leq l \leq l_j$$

5.6 Artificial Neural Networks

Artificial neural networks are computing systems inspired by the biological neural networks of animal brains. Based on a collection of connected units or nodes called artificial neurons, it loosely model the neurons in a biological brain. Each connection, or edge, can transmit a signal to other neurons. An artificial neuron receives signals then processes them and can signal neurons connected to it. The "signal" at a connection is a real number and the output of each neuron is computed by a non-linear function of the sum of its inputs. Neurons and edges typically have a weight which adjusts as learning proceeds, increasing or decreasing the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses the threshold. Typically, neurons are aggregated into layers and different layers may perform different transformations on their inputs. Signals travel from the first layer to the last layer, possibly after traversing the layers multiple times.

NNs learn by processing examples and forming probability-weighted associations between the two, which are stored within the data structure of the net itself. The training is usually done by determining the difference between the processed output of the network and a target output. This difference is the error. The network then adjusts its weighted associations according to a learning rule and using this error value. Successive adjustments will cause the NN to produce output increasingly similar to the target output. After a sufficient number of these adjustments, the training is terminated based upon a predetermined criterion.

5.6.1 Bayesian Regularized Neural Networks

The final method we will use to estimate a prediction model is Bayesian regularized neural network. It is more robust than standard back-propagation nets and can reduce or eliminate the need for lengthy cross-validation. Bayesian regularization is a mathematical process which converts non-linear regression into a statistical problem in the form of ridge regression. It is both difficult to overtrain, since evidence procedures provide an objective Bayesian criterion for stopping training, and difficult to overfit, since it calculates and trains on a number of effective network parameters or weights, effectively turning off those which are not relevant. Automatic relevance determination can be used to "estimate" the importance of each input. Bayesian Regularized NNs can be presented through the following equation:

$$F = \beta E_D + \alpha E_w$$

Where:

$$E_D = sum\ of\ squared\ errors$$

$$E_W = sum\ of\ the\ squared\ network\ weights$$

The function term seeks to (E_D) and (E_W) while maximizing the objective functions beta and alpha. In the Bayesian network, the weights are considered random variables and thus their density function is written according to Bayes' rules:

$$P(w|D, \alpha, B, M) = \frac{P(D|w, \beta, M) * P(w|\alpha, M)}{P(D|\alpha, \beta, M)}$$

$$w = vector\ of\ network\ weights$$

$$D = data\ vector$$

$$M = neural\ network\ model$$

Assuming noise is Gaussian, the probability density function for the weights is computed. The optimization of the regularization parameters alpha and beta require solving the Hessian matrix of F(w) at the minimum point w^{MP}. Foresee and Hagan (Foresee & Hagan, 1997) proposed a Gauss-Newton approximation to the matrix which is possible if the Levenberg-Marquardt training algorithm is used to locate the minimum. This technique reduces the potential for arriving at local minima, thus increasing the generalizability of the network. Bayesian regularized networks also penalize overly complex models as unnecessary linkage weights are effectively driven to zero. Finally, the network trains on the remaining nontrivial weights.

6. Results

6.1 Models' Precision

We will begin our analysis of results by evaluating the precision of our optimal models which were derived by applying the methods described in the previous section on data described in the Data section. To evaluate the accuracy of the predictions by the models, we will use the mean squared error (MSE), as this is a conventionally accepted approach to evaluating the goodness-of-fit of quantitative models in research. Table 3 presents the MSE of each model calculated for predictions made on the training set as well as on the test set. As a supplement, Figure 7 illustrates the predicted fund returns in comparison to the actual observations of fund returns. Ultimately, the perfect model would have each value of the predicted returns equal to the value of actual observed returns, on a diagonal line of the form of $f(x) = x$.

According to the bias-variance trade-off described in section 3.2, we expect that the most flexible models, i.e., SVM and NN, will have the lowest training MSE, as the more flexible models have stronger potential for discovering and replicating patterns in the data. Correspondingly, we expect the less advanced models, i.e., the Linear model and Ridge, to have a smaller difference between the test MSE and the training MSE. This is because the more flexible advanced models will most probably include noise in their predictions while calculating their algorithm, and hence they will overfit the algorithm in the training set, resulting in a higher MSE for the test set.

Model	Training MSE	Test MSE
Linear	0.0442	0.0370
Ridge	0.0588	0.0387
SVM	0.0072	0.0538
NN	0.0423	0.0377

Table 3: Predictor Models' MSE

As seen from Table 3, The Linear model, Ridge, and the NN model display similar MSE for the training set. Meanwhile, the Support Vector Regression model outperforms the other models in terms of training MSE, which is most probably due to the Support Vector Regression having the most flexibility out of the set of methodologies we use. However, the model's test MSE is performing notably worse than its training MSE, which can be further seen by the larger spread between the model's predicted NIRR and the actual observed NIRR in Figure 7. This most likely implies that the model overfit the data while calculating the algorithm for the test set, hence resulting in poorer performance on out-of-sample data. The latter is coherent with the expectations on ML theory were previously outlined in Table 1.

The best-performing model on out-of-sample data is the Linear model, as seen by its test MSE, which is again coherent with our expectations described in the second paragraph of the current section of the paper – i.e., the less advanced models performing stronger on the test set due to not overfitting the data when calculating the algorithm for the training set. We also see that both Ridge and the Neural Network methods produce approximately similar results in the test set, with only a marginal difference to that of the Linear model. NN performing stronger than SVM on the test set and giving a poorer performance than SVM on the training set has a possible explanation in the NN methodology including an objective Bayesian criterion which aims at preventing overfitting the model by stopping to train the model when there is a chance of the former, as discussed in the Methodology section. Meanwhile, the Linear model, Ridge and the NN model performing well on the test set to the extent where their test set MSE is lower than their training set MSE is somewhat surprising. From Figure 7 we can see that the training set includes a notably larger number of extreme observations of NIRR above 1 or below -0.5, which is likely due to chance from when the observations were randomly divided into the training set and the test set. Therefore, a possible explanation for the Linear model, Ridge and the NN model performing stronger on the test set than on the training set involves the former containing less variation of observations than the latter. SVM not benefiting from the lower variation in the test set is, as discussed, likely due to the model overfitting the data when calculating the algorithm for the training set.

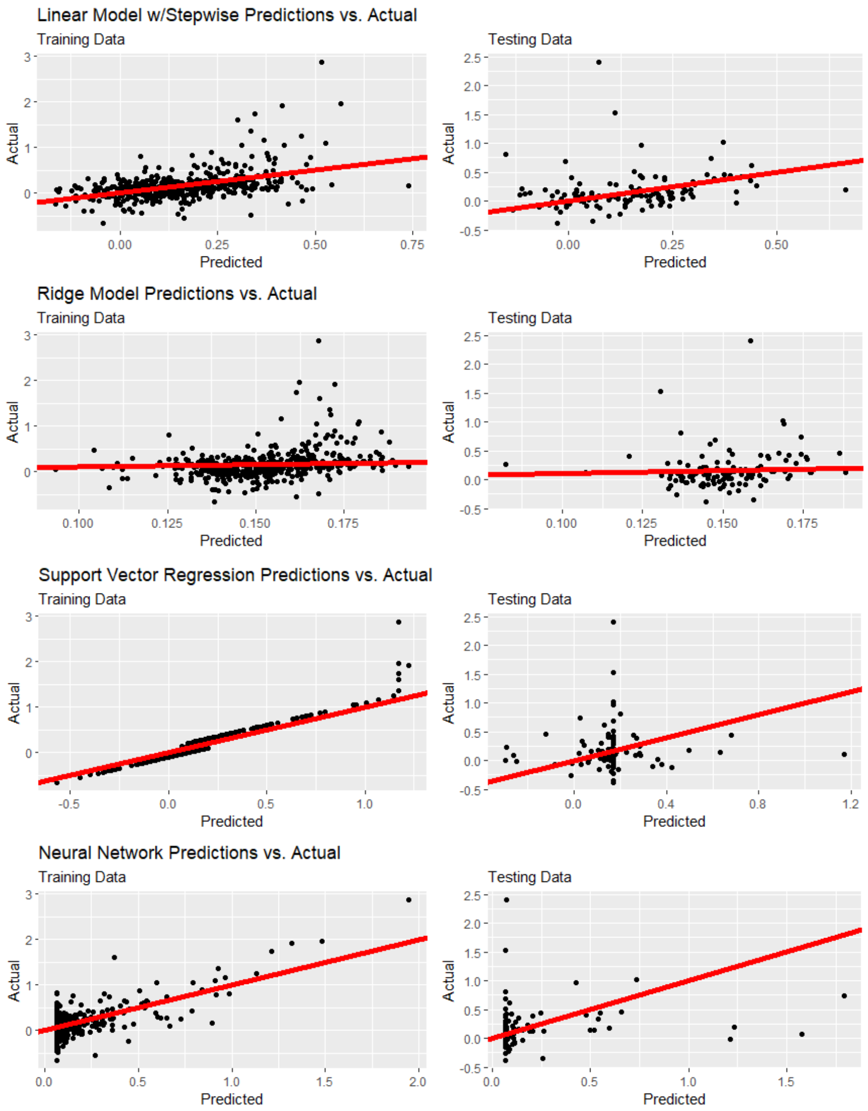


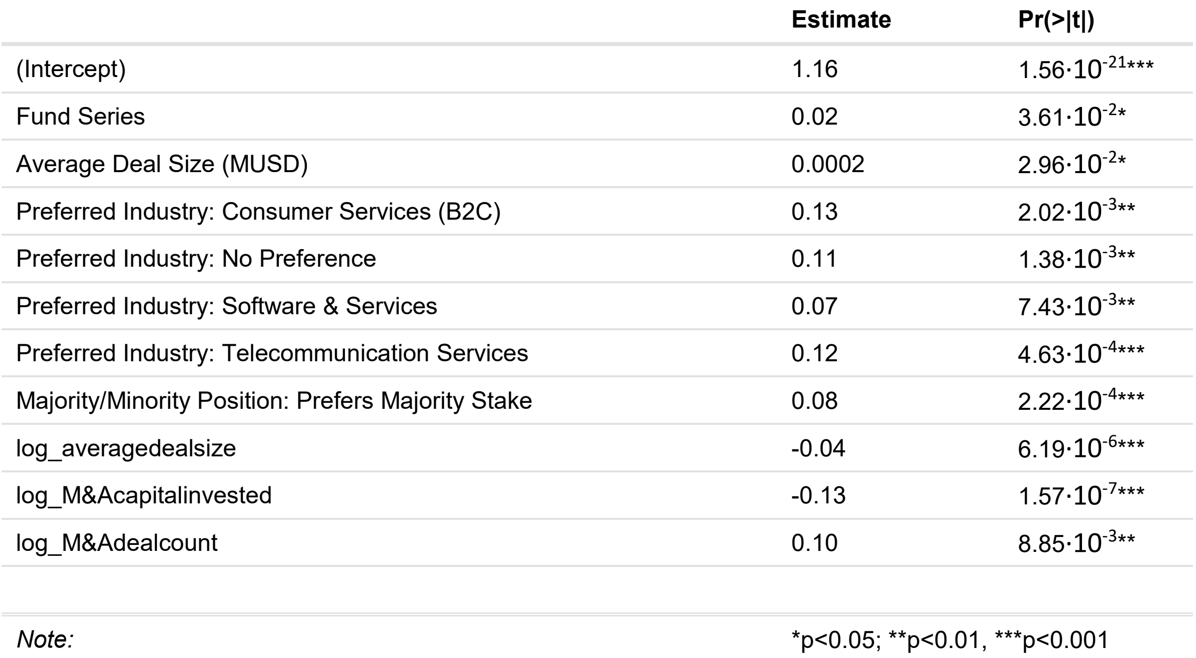
Figure 7: Predicted NIRR compared to Actual NIRR

6.2 Analysis of Coefficients

In this section, we will evaluate coefficient of variables appearing in our final models, in order to understand what drives the models' predictive power. We will do so by comparing coefficient to previous research and our own hypotheses described in the Data section.

6.2.1 Linear Model's Coefficients

As the Linear model produces the strongest results in terms of model precision on the test set, we will begin by examining the coefficients of the Linear model. Table 4 illustrates the coefficient appearing in the abovementioned model which are statistically significant at a 5% level. While evaluating the results, we have also examined the coefficients of categorical variables to ensure these are not driven by outliers inside each category, coming to the conclusion that this is not the case.



	Estimate	Pr(>|t|)
(Intercept)	1.16	$1.56\cdot10^{-21}$***
Fund Series	0.02	$3.61\cdot10^{-2}$*
Average Deal Size (MUSD)	0.0002	$2.96\cdot10^{-2}$*
Preferred Industry: Consumer Services (B2C)	0.13	$2.02\cdot10^{-3}$**
Preferred Industry: No Preference	0.11	$1.38\cdot10^{-3}$**
Preferred Industry: Software & Services	0.07	$7.43\cdot10^{-3}$**
Preferred Industry: Telecommunication Services	0.12	$4.63\cdot10^{-4}$***
Majority/Minority Position: Prefers Majority Stake	0.08	$2.22\cdot10^{-4}$***
log_averagedealsize	-0.04	$6.19\cdot10^{-6}$***
log_M&Acapitalinvested	-0.13	$1.57\cdot10^{-7}$***
log_M&Adealcount	0.10	$8.85\cdot10^{-3}$**
Note:	*p<0.05; **p<0.01, ***p<0.001	

Table 4: Statistically Significant Coefficients from Linear Model

Fund Management's Experience:

As seen from Table 4, in our final Linear model, "Fund Series" has a positive coefficient of 0.02, meaning all else kept equal, increasing the sequence of the fund series yields an additional 2% NIRR. This is in line with our expectations and previous research of fund series sequence and fund managements' experience having a positive relationship with fund returns. However, unlike Roggi (2019) we do not observe a concave effect of fund series on NIRR, and only a linear one, as there are no logarithmic fund management variables turning significant in our model. This suggests the contribution of fund management's experience towards returns increases linearly with each fund series, and is significant for more than high fund series sequence numbers.

We also do not observe the variable "Fund Number" as part of the statistically significant variables in our model. A possible explanation is that fund managers launching funds from unrelated fund series which have a different fund strategy does not allow for any synergy effects with the existing funds, thereby no learning effects occur, and the funds' returns are not affected. Finally, the variable "Number of Partners" does not appear statistically significant in our model. A possible reasoning for this could be differences in how this variable is reported among different funds – while some funds might include employees outside of the senior management team in this variable for advertisement purposes, other funds, for reasons of protecting their inside information, could include only part of their senior management team.

Additionally, a possible reasoning could be differences in rules for becoming a partner among different funds. A possibly more precise approach to measuring the experience and extensiveness of the network of the fund managements' team could be through variables "Average Years of Experience" and "Total Deals Previously Completed", as these allow for fewer reporting differences and unlike "Number of Partners" are not affected by possible different rules for becoming a partner. These variables were not used as proxies in our analysis due to unavailability of data on the former.

Funds' Strategy:

Variable "Average Deal Size" is statistically significant in our model, with a small coefficient of 0.0002, while the logarithm of the variable has a negative coefficient of -0.04. With the coefficient of a logarithm of a variable showing the effect on the output variable with a percentage change in the predictor variable, the results imply that at low values of average deal size, an increase of 1 MUSD will equal to more than a percentage increase, yielding a decrease in NIRR from increasing average deal size. Meanwhile at higher values of deal size, an increase of 1 MUSD will equal less than a percentage increase, yielding an increasing in NIRR with increased deal size. This fits well with our expectations and Loos' (2006) findings of investments in small- and mega-cap categories displaying the highest returns for their corresponding funds. We also do not observe variable "Fund Size" in our model, which suggests a driving force behind private equity fund returns is the average size of the companies they invest in, as it is these companies the sale of which will yield a cash inflow to the fund, and as long as the fund managers are able to close the fund, the size of the fund is not equally significant.

Further, we observe a positive coefficient for preference for industries Consumer Services (B2C), Software & Services, and Telecommunication Services. This stands in support of the resource based theory described in the Data section, where specialization for an industry is seen as enabling the funds to build a competitive advantage through increased understanding of their focus area, allowing the funds to support their portfolio companies more effectively and produce higher returns. Nevertheless, having no preference for which industry or set of industries the funds will invest in having a positive coefficient is strange, as it supports the opposing view of the portfolio theory, where it is argued that when diversification is reduced, the number of potentially profitable investments is reduced for the fund as well, leading to weaker performance. When examining the statistically significant coefficients further, we can see that a number of preferred industries are not statistically significant, which is likely a result of having few observations for preference for these industries. Therefore, it is possible that the coefficient of having no preference for an industry or set of industries is positive due to it including the effect of a specialist preference for each of the industries for which we have too few observations for them to be statistically significant.

Additionally, our model shows the preference for a majority stake to have a positive coefficient. This supports our expectations and Loos' (2006) findings, which are built on the rationale that majority stakes allow more experienced fund managers to apply their experience more effectively in the management of their portfolio companies, hence leading to stronger strategic improvement of their companies and higher returns.

We can also notice the absence of a few Funds' Strategy variables, the first of which being the preference for the geographic region where the funds will conduct their investments. These results stands in line with our expectations, as we initially believed the impact of geographic preference could potentially be explained by other factors, such as the macroeconomic environment of- or the PE market conditions of the geographic region for investment, as these variables can capture the majority of the effects of preferring a geographic region for investment. Next, the "Fund Type" variable does not appear significant in our model either. The most likely explanation involves that the main characteristics of the variable are already captured in the "Average Deal Size" and "Majority/Minority Position", which are statistically significant, and splitting up the variable "Fund Type" into its characteristics explains the variation in the output variable better than pooling these characteristics together. Finally, we see variable "ESG Focus" not turning significant, which can be expected due to the very limited number of observations of funds with an ESG focused strategy in our data sample.

PE Market Variables:

From Table 4 we can observe two PE market variables, the logarithm of M&A volume's deal count, and the logarithm of M&A volume's capital invested, which are essentially measures of the same factor, showing opposing coefficients, with deal count being positive and capital invested being negative. These results suggest that a favourable market environment for deals is highly important for private equity firms to find promising deals and later exit these deals, as seen through a positive coefficient for deal count. Simultaneously, increasing the capital invested in M&A transactions can drive up valuations, making it later difficult to find promising deals at affordable values and hence reducing returns for funds.

Variables “EV/Revenue” and “EV/EBITDA” do not appear statistically significant in our model. A part of the possible explanation is based on the reasoning above, i.e., that the effect of valuations on fund returns can be captured through the volume of capital invested in M&A transactions. While the variables “EV/Revenue” and “EV/EBITDA” break this effect down into two measures of valuation, there is a general difference in which of these measures are used for transactions based on the type of fund. As discussed in the Theory section, “EV/Revenue” is most applied to early stage companies which Venture funds invest in, while “EV/EBITDA” is most applicable to profit-making later staged companies in which Buyout funds invest in. As each of the types of funds rarely uses the other measure, capital invested in M&A transactions might be better suited to capture the effect of valuations on fund returns.

Furthermore, variable “Dry Powder Market” is not statistically significant in our model, which most probably suggest that M&A volume is a more precise proxy for the effect of the conditions of the dealmaking market on private equity returns. Finally, the variable “Funds Raised Market” being absent from our set of statistically significant variables is strange, as prior research mentioned in the Data section has shown the importance of successful fundraising for private equity fund returns. However, this stand in line with our evaluation of results for Funds’ Strategy variables, namely that raising a fund at a specific size is less important than the average size of the companies the fund will invest in, as it is the sale of these companies that is the factor giving inflow of cash to the fund.

Macroeconomic Variables:

Interestingly, none of the macroeconomic variables turn out statistically significant in our model, which goes against our expectations and previous research discussed in the Data section. A possible reasoning for this result is the inclusion of PE market variables in the model, as these essentially explain the same market effect on each fund as macroeconomic variables do – as evident from Appendix A.2, macroeconomic variables are highly correlated with PE market variables. Moreover, in the case of macroeconomic variables, using data aggregated on a regional level might give inaccurate results as a comparison of the difference in data when going from region to country has led us to draw the conclusion that such difference is notably larger for macroeconomic variables than for PE market variables. This data was not collected on a country level for macroeconomic variables as historical records of such data are extremely limited in their year-range for some countries, and therefore, basing the macroeconomic data on a country level would severely limit the size of our dataset, which we hence opted not to do.

6.2.2 Advanced Models' Coefficients

Further, we will examine the coefficients of the model which produces the strongest results in terms of test set MSE, i.e., the SVM model. The coefficients of models which are not optimal in neither test set nor training set MSE are presented in Appendix B.3. While coefficients in linear models can be interpreted as the direct effect on the output variable from a unit increase in the predictor variable, each coefficient in ML models includes interaction effects with other predictor variables. Therefore, there is a lack of general consensus among researchers on the approach for interpreting ML models' coefficients. A much cited paper in this area by Guyon (2002) uses the square of the coefficients from SVM models as a ranking of relative importance of the corresponding variables for the models' predictive power. As such, this will be the approach we will be applying in this section.

Variable	Squared coef.
M&A Volume (Capital invested, MUSD)	$1.2 \cdot 10^{14}$
M&A Volume (Deal count)	$9.5 \cdot 10^{9}$
Dry Powder Total (MUSD)	$3.3 \cdot 10^{9}$
Funds Raised Total	$2.3 \cdot 10^{7}$
Fund Size (MUSD)	$6.6 \cdot 10^{6}$
Average Deal Size (MUSD)	$4.0 \cdot 10^{4}$
log_M&Acapitalinvested	1,691
log_totaldrypowder	1,614
EV/Revenue	1,057
EV/EBITDA	1,013
log_M&Adealcount	739
log_fundsraisedtotal	521
Number of Partners	152
Fund Series	112
log_fundsize	86
log_dealsize	53
Preferred Geography: No preference	15
Preferred Industry: Pharmaceuticals, Biotechnology & Life Sciences	14
Preferred Industry: Healthcare Equipment & Services	11
Preferred Industry: Telecommunication Services	9
Fund Number	7
Majority/minority stake: Prefers majority stake	5
log_fundnumber	5
Preferred Industry: Software & Services	4
log_fundseries	4

Table 5: Squared Coefficient of 25 most significant variables in SVM Model

Table 5 presents the squared coefficients from our optimal SVM model ranked after significance for the model's predictive power. As evident from the table, the SVM model contains a convex relationship between M&A Volume and NIRR through the logarithm of M&A Volume measured in both capital invested and deal count. This is consistent with the results from our Linear model discussed in the previous section. However, in contrast to the Linear model, the non-logarithmic M&A Volume, through both capital invested and deal count, is the most vital driver of the SVM model. As previously discussed, the main difference between the SVM model and the Linear model includes the SVM model taking into account interaction effects between predictor variables. Therefore, the most likely explanation for the non-logarithmic M&A Volume being among the most significant variables in the SVM model but not in the Linear model is the close correlation between this variable and the logarithm of it – as the latter already appeared significant in the Linear model. This is confirmed in the Variable Correlation Matrix in Appendix A.2 which, as expected, shows a strong positive correlation between M&A Volume and the logarithm of the former. For its part, this creates an interaction effect between M&A Volume and the logarithm of M&A Volume, where an increase in the former leads to an increase in its logarithm. As discussed in subsection 6.2.1, the logarithm of M&A volume measured in deal count has a positive effect on funds' returns at liquidation, while logarithm of M&A volume measured in capital invested has a negative effect on funds' returns. This yields two interaction effects, where an increase in M&A Volume in deal count increases its logarithm, decreasing fund returns, while an increase in M&A Volume in capital invested increases the corresponding logarithm, boosting returns.

In further contrast to the Linear model however, we see that the PE market's "Dry Powder Market" is the third most significant driver of the SVM model's predictive power. The most likely explanation for "Dry Powder Market" being among the most significant variables in the SVM model but not in the Linear model is that as private equity funds exit their investments and thereby increase their cash at hand which they can reinvest, i.e., their Dry Powder, they largely do so by exiting their investments through M&A transactions, hence increasing M&A volume. This creates a positive relationship between "Dry Powder Market" and M&A Volume, which is confirmed in Appendix A.2. As mentioned in the previous paragraph, the logarithm of M&A Volume measured in deal count has a positive effect on funds' returns at liquidation, while logarithm of M&A volume measured in capital invested has a negative effect on funds' returns. Therefore, while the interaction effect between "Dry Powder Market" and the logarithm of M&A Volume measured in both capital invested and deal count implies that an increase in “Dry Powder Market” gives an increase in the logarithm of M&A Volume, it is difficult to determine what the final impact of this interaction effect is on funds returns. We also see that the logarithm of Dry Powder Market is among the most significant variables in the model, although not as significant as the “Dry Powder Market” itself. This confirms the convex relationship between “Dry Powder Market” and fund returns, which we lined out as our expectation in the Data section. Appendix A.2 illustrates a positive relationship between the logarithm of M&A Volume and the logarithm of Dry Powder Market, and hence we can expect that an increase in the logarithm of Dry Powder Market yields an increase in the logarithm of M&A Volume, both in terms of capital invested and deal count. While the logarithm of M&A Volume in capital invested has a negative effect on fund returns in the Linear model, the logarithm of M&A volume in terms of deal count has the opposite effect. Therefore, it is again difficult to say what the total effect of the logarithm of Dry Powder Market on fund returns at liquidation is.

Next, the PE market variable “Funds Raised Market”, which was not significant in the Linear model, appears highly significant in the SVM model. A possible rationale behind this is that as private equity funds raise a larger number of funds, these funds begin their investments phase, purchasing companies through largely M&A transactions and thereby increasing M&A Volume. This creates a positive interaction effect between the variable “Funds Raised Market” and M&A Volume, and we can see a confirmation of such interaction effect in the Variable Correlation Matrix in Appendix A.2. As discussed in the previous paragraph, the Linear model containing both the logarithm of M&A Volume measure in deal count and in capital invested, with each of the two variables having an opposite effect on NIRR. This again makes it difficult to say what the final effect of the interaction effect between variable “Funds Raised Market” and M&A Volume has on NIRR. Furthermore, we also see a somewhat less significant convex relationship between “Funds Raised Total” and funds returns through the logarithm of Funds Raised Market, which stands in line with our expectations from the Data section. As Appendix A.2 shows a positive relationship between the logarithm of M&A Volume and the logarithm of Funds Raised Market, an increase in the logarithm of Funds Raised Market is expected to have a positive effect on the logarithm of M&A Volume. For a corresponding reason to that described in the previous paragraph, it becomes difficult to estimate what the total effect of the logarithm of Funds Raised Market is on fund returns at liquidation.

Furthermore, while both the Linear and the SVM model contain variable “Average Deal Size” as a significant driver of the models’ predictive power, the SVM model also includes “Fund Size” as a significant variable. A potential rationale behind this involves that due to the difficulty in managing a large number of deals, large funds could prefer to invest their capital in investments with a large Average Deal Size. Appendix A.2. confirms such strongly positive relationship between variables “Average Deal Size” and “Fund Size”. An increase in the variable “Funds Size” can therefore be expected to yield a positive interaction effect on “Average Deal Size”, and as the variable “Average Deal Size” has a positive effect on NIRR in the Linear model, increasing “Fund Size” results in further boosting NIRR. Additionally, we find that “Fund Size” has a concave relationship with fund returns through the logarithm of Fund Size. From Appendix A.2, we see that the logarithm of Fund Size is positively correlated with the logarithm of Average Deal Size. An increase in the logarithm of Fund Size can therefore be expected to have a positive effect on “Average Deal Size”, and as the logarithm of “Average Deal Size” has a negative effect on NIRR in the Linear model, increasing the logarithm of “Fund Size” results in reduced fund returns.

Valuation multiples EV/Revenue and EV/EBITDA are another instance of variables which appear as significant drivers of the model’s predictive power for the SVM model, despite not being statistically significant in the Linear model. This difference can potentially be explained in that as valuation multiples increase, this spurs an increase in M&A transactions due to private equity funds exiting their investments in favourable market conditions. Accordingly, a decrease in valuation multiples can result in an increase in M&A transactions as private equity funds increase their purchases of privately held companies due to being able to purchase these companies at low valuations. Supporting evidence of this can be found in Appendix A.2, with EV/Revenue and EV/EBITDA being highly positively correlated with the logarithm of M&A Volume measured in capital invested. This yields an interaction effect between valuation multiples and the logarithm of M&A Volume in capital invested. As the logarithm of M&A Volume in capital invested has a negative effect on NIRR, we can expect the interaction effect to be negative, with an increase in valuation multiples resulting in an increase in the logarithm of M&A Volume in capital invested, which for its part reduces fund returns.

While the Linear model only included “Fund Series” as a significant variable for the model from the set of predictor fund management variables, the SVM model contains additional variables through “Number of Partners” and “Fund Number” from the mentioned set of predictor variables. The rationale behind the former involves the following: upon examining the attributes of the variable “Average Deal Size”, we observed that fund which conduct large sized deals on average, often undergo somewhat fewer deals than funds conducting smaller sized deals. Therefore, in order to manage the larger number of ongoing deals, fund with smaller sized deals could be recruiting a larger number of partners. Evidence of such negative relationship between variables “Number of Partners” and the logarithm of Average Deal Size can be found in Appendix A.2. As the logarithm of Average Deal Size has a negative effect on NIRR, this yields a positive interaction effect, with an increase in “Number of Partners” giving a decrease in the logarithm of Average Deal Size, which for its part yields increases the fund returns. Meanwhile, variable “Fund Number” is likely more statistically significant in the SVM model than in the Linear model due to one of its components being variable “Fund Series”, which was already found to be statistically significant in the Linear model. As seen from Appendix A.2, “Fund Number” is strongly correlated with variable “Fund Series”. As variable “Fund Series” has a positive effect on NIRR in the Linear model, this gives a positive interaction effect where an increase in “Fund Number” often comes with an increase in “Fund Series”, thus boosting fund returns. However, as “Fund Series” is a component of “Fund Number”, an increase in the latter does not necessarily imply an increase in “Fund Series”. This is likely the reason for the variable “Fund Number” being placed quite low in the ranking of variable importance for the model’s predictive power. We also see that both the variable “Fund Series” and “Fund Number” have an additional parabolic effect on NIRR through the logarithms of the mentioned predictor variables, which is in line with our expectations from the Data section.

The final notable difference between the Linear and SVM model involves the latter including geographic specialization variables as well as different set of industry specialization variables than that of the Linear model. The most probable explanation for said difference is that the effect of geographic and industry specialization on NIRR is dependent on an interaction effect between each of the two types of specialization and the fund manager’s experience, measured by Fund Number, Fund Series, and Number of Partners in our analysis. Through this, experienced managers might have developed a competitive advantage through increased understanding of their investment area, and therefore, specialization allows them to utilize said experience to achieve stronger returns. This stands in support of the resource based theory described in the Data section. Meanwhile, less experienced managers might lack an intricate understanding of specialized investments into a specific industry or geographic region, and could therefore deliver higher returns by diversifying their investment area. This rationale is backed by the portfolio theory which was lined out in the Data section. Appendix A.2 only gives an overview of correlation between numerical variables, as the large number of categories in categorical variables in the dataset would make the Variable Correlation Matrix unobservable if all said variables were included. This makes it quite difficult to examine whether the abovementioned rationale is in line with the data. However, we can find an indicator of said rationale being correct in the type of specialization variables included in the model, particularly in the inclusion of industry specialization for Biotechnology and Healthcare Equipment & Service. These are notably high-risk industries requiring extensive experience and deep knowledge of the industry to succeed. Additionally, we can see such an indicator through the inclusion of geography specialization for no preferred geography, i.e., a preference for diversifying investments among multiple geographic regions, which could fit well for less experienced fund managers who lack intricate understanding of a specific geographic or industry category. This suggests that the SVM model including a higher number of as well as different types of industry and geographic specification variables than the Linear model is indeed largely due to interaction effects between said predictor variables and fund management experience variables.

7. Conclusion

7.1 Conclusion

Our research had the objective of using machine learning methodologies to predict the performance of private equity funds, in order to allow for more effective fund selection for potential retail investors entering the private market. To accomplish this, we began by conducting a Linear AIC regression analysis, in order to produce the set of variables which best explain the variation in the return measure. Next, we have employed more advanced ML algorithms, which take into account possible interaction effects between variables to compute the model which best predicts the return measure. We selected NIRR as our measure of return, due to both it being the return measure of choice in the private equity industry, as well as it being notably similar to the annualized returns measure used in public markets, simplifying its interpretability for retail investors at who this paper is primarily directed. Further, we divided the predictor variables for the analysis in four groups of variables – fund managements' experience, funds' strategy, PE market variables, and macroeconomic variables.

Our optimal Linear AIC model performs the strongest in terms of precision on the out-of-sample test set, delivering an MSE value of 0.0370. This was in line with our expectations, as the more advanced models tend to overfit the data when calculating the optimal algorithm for the training set, resulting in a poorer performance on the out-of-sample test set. Meanwhile, the more advanced ML models, in particularly SVM, displays the strongest precision on the training set, with an MSE value of 0.0072. This was again coherent with our expectations, as the more advanced models allow for more flexibility by incorporating potential interaction effects between the predictor variables, hence yielding a stronger precision on the test set.

An analysis of the statistically significant coefficients in the Linear regression showed that the model predicts NIRR through a positive coefficient for the funds' sequence in their respective Fund Series, a convex effect between the Average Deal Size and NIRR, a stronger NIRR for funds preferring majority stakes in their deals, a negative coefficient for the logarithm of M&A Volume measured in capital invested at fund inception year, and a positive coefficient for the logarithm of M&A Volume measured in deal count at fund inception year. Additionally, the model assigns a stronger return for funds preferring to conduct deals in industries involving Consumer Services (B2C), Software & Services, Telecommunication Services, and funds diversifying across multiple industries.

Upon adding flexibility by incorporating potential interaction effects between predictor variables through the SVM model, additional variables became statistically significant. An analysis of the coefficients of such variables showed that in addition to the variables which are statistically significant for the Linear model, the SVM model's predictive power is also driven by the non-logarithmic M&A Volume, measured in both capital invested and deal count, the PE market's "Dry Powder Market" and "Funds Raised Market" at the fund inception year, the PE market's EV/Revenue and EV/EBITDA at fund inception year, the funds' "Number of Partners", the funds' sequence in their respective "Fund Number" from the same fund manager, as well as the logarithm of said variables, which yields a parabolic effect between such variables and NIRR. Additionally, when adding flexibility through the SVM model, the set of preferences for industries and geographical regions to which the Linear model assigned a higher return saw some changes. For the advanced SVM model, this set of variables includes a preference for industries involving Pharmaceuticals, Biotechnology & Life Sciences, Healthcare Equipment & Services, Telecommunication Services, and Software & Services, as well as funds having no preference for geographic region for investment, i.e., a preference for diversifying investments among multiple geographic regions. As discussed in detail in the Results section, the likely reason for this different set of Preferred Industry and Preferred Geography variables involves possible interaction effects between these variables and fund management's experience variable. This involves the mentioned industries being high-risk industries where fund managers who have developed extensive experience in said industries can have a competitive advantage, allowing them to deliver higher returns by specializing in these industries, while less experienced fund managers could instead deliver higher returns by diversifying their investments across multiple geographic regions.

7.2 Biases, Limitations & Future Research

In this section we will address the biases and limitations which can affect our results, as well as discussing recommendations for future research which can improve our study. A first bias which has potential for impacting our results is the selection bias, which involves funds which perform poorly not being interested in reporting their returns due to fear of it having a negative effect on future fundraising. A similar bias is the survivorship bias, where managers who discontinue their private equity firms due to poor previous fund performance do not report the returns from their funds, yielding a sample of only successful funds which have "survived" the initial stages. A possible effect of these two biases on our results includes our models potentially overestimating NIRR due to our dataset containing too few funds with poor returns. A potential solution could be to directly contact the funds with poor returns to request their data. However, this is quite difficult to accomplish in practice as it requires intricate knowledge of which funds achieved poor returns without these returns being reported, as well as due to the voluntary basis on which fund managers may choose to share or not to share their data.

A further bias which can impact our results is the omitted variables bias, which involves potentially statistically significant variables not being included in the models, hence somewhat skewing the estimates for the variables which the models contain. One such omitted variable could include the funds' past performance, as one can expect managers who have shown strong returns in the past to have developed a competitive advantage through a deeper understanding of their investment area. Our dataset is quite limited by a total of 600 funds, with a large number of these funds not having reported any previous funds from the same fund series or the same fund manager. Therefore, including a measure of past performance would significantly reduce the size of our dataset, potentially turning some of the statistically significant variables in our models into statistically insignificant. As such, we have chosen to not include a measure of past performance in our analysis. Another omitted variable could be a more representative proxy for management experience, which could be the average number of years of experience for the fund management's team, or the total number of deals previously conducted by the fund management. The rationale for this is that unlike variables "Fund Number" and "Fund Series", the previously mentioned variables take into account the different number of deals conducted by the management for each of their previous funds, as well as the relevant working experience the management team had before launching their first fund. These variables were not included in our analysis due to such data being inaccessible to us, and it would be interesting to see if adding these variables would strongly improve our models' precision.

Next, as discussed in section 6.2.1, we collected macroeconomic data on a regional level, due to historical records of such data being extremely limited in their year-range for some countries. Having collected macroeconomic data on regional level might be the reasoning for said variables appearing statistically insignificant in our models, as a closer examination made us draw the conclusion that the difference in data between region and country level is quite large for macroeconomic variables. Therefore, being able to collect the macroeconomic data on country level could be a further possible improvement for future research on this study. Additionally, as described in the Theory section, using NIRR as a measure of returns has its limitations through the measure's underlying assumption that distributions to investors are reinvested at the same rate of returns along the time horizon, which is not always realistic. We have not adjusted for this issue due to our data sample only providing raw NIRR data and lacking information on size and timing of distributions. Therefore, a further improvement in future research would include adjusting NIRR such that distributions to the investors are reinvested at a different rate, for instance at the rate of the private equity firm's cost of capital.

A final limitation is related to the size of our dataset when applied in machine learning algorithms. ML models are attractive as they can learn and build more complex models which offer more flexibility and often yield better precision in output variable prediction. This is due to the many combinations of interaction effects between input variables which machine learning algorithms can choose from when constructing their model. Subsequently, these algorithms need to estimate a large number of coefficients. In case if the dataset is small and some of the variables only have a few observations, this leads to ML models starting to memorize the pattern between these variables in the training set instead of generalizing it. This leads to overfitting the prediction model, and hence less precise results when applying the model to the testing set. A rule of thumb for determining whether a dataset is small for ML methods is to divide the number of observations by the number of features – if the number is less than 10, the dataset is seen as small. Our dataset has approximately 8.6 observations per feature, meaning it should be considered a small dataset.

In order to address this issue, we conduct the Linear model analysis before using the more advanced methods, as it enables the removal of variables which we find to not have a statistically significant relationship with the output variable. Further, we divide the dataset into a training set and a test set of correspondingly 480 and 120 observations, in order to analyze whether there is overfitting in our prediction models. While using the advanced ML methods, we utilize the methods' benefits in order to reduce overfitting. For Support Vector Regression this involves the method's aims to minimize error made by the learning machine on out-of-sample data, which gives good generalization capability and thus, it tends to perform well on data outside the training set. Meanwhile, for a Bayesian Regularized Neural Network, the likelihood of overfitting is reduced as it utilizes evidence procedures that provide an objective Bayesian criterion for stopping training when there is a chance of overfitting. Additionally, the method trains on several network parameters, effectively turning off those which are not relevant. Nevertheless, increasing the size of the data sample could be a factor towards improving the precision of our results.

Bibliography

Achleitner, A.-K., Braun, R., Engel, N., Figge, C., & Tappeiner, F. (2010). Value Creation Drivers in Private Equity Buyouts: Empirical Evidence from Europe. *The Journal of Private Equity*.

Aigner, P., Albrecht, S., Beyschlag, G., Friederich, T., Kalepky, M., & Zagst, R. (2008). What drives PE? Analyses of success factors for private equity funds. *Journal of Private Equity, 11*(4).

Alon-Beck, A. (2020, November 24). Democratizing Access To Private Markets With The Rise Of Alternative Venture Capital.

Brown, G. W., Harris, R. S., Tim, J., Kaplan, S. K., & Robinson, D. (2015). *What Do Different Commercial Data Sets Tell.* Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2701317

Cambridge Associate LLC. (2022, Q2). *U.S. Private Equity Benchmarks (Legacy Definition) Q2 2020 Final Report*. Retrieved from Cambridge Associates: https://www.cambridgeassociates.com/private-investment-benchmarks/

CB Insights. (2022, October 13). $1B+ Market Map: The world's 1,191 unicorn companies in one infographic.

CFA Institute. (2022, December 8). *What Does a Career in Private Equity Look Like?* Retrieved from CFA Institute Website: https://www.cfainstitute.org/en/programs/cfa/charterholder-careers/roles/private-equity

Christensen, J. L. (2007). The Development of Geographical Specialization of Venture Capital. *European Planning Studies*.

Comittee on Capital Markets Regulation. (2019). *Expanding Opportunities for Investors and Retirees: Private Equity.*

Corporate Finance Institute. (2022, December 10). *Data Science & Machine Learning Fundamentals*. Retrieved from Corporate Finance Institute Website: https://corporatefinanceinstitute.com/course/data-science-and-machine-learning/

Dalpiaz, D. (2018). Bias-Variance Tradeof. *STAT 430.* Urbana-Champaign: University of Illinois at Urbana-Champaign.

Duong, J. (2014). *The Risks, the Returns and the Fundraising Successes of Private Equity Funds.* Preqin.

Easton, P. D., Larocque, S., & Stevens, J. S. (2021). Private Equity Valuation Before and After ASC 820. *SSRN Electronic Journal*.

Esty, D. C., & Cort, T. (2020). *Private Equity and ESG Investing.* Palgrave Macmillan.

Finley, J. (2019, November). Expanding Retail Access to Private Markets.

Foresee, F. D., & Hagan, M. (1997). Gauss-Newton Approximation to Bayesian Learning. *International Conference on Neural Networks.* Houston, TX, USA.

Gupta, V., Koller, T., & Stumpner, P. (2021, October 21). Reports of corporates' demise have been greatly exaggerated.

Guyon, I., Weston, J., & Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*(1-3).

Harris, R. S., Jenkinson, T., Kaplan, S. N., & Stucke, R. (2022). Has Persistence Persisted in Private Equity? Evidence from Buyout and Venture Capital Funds. *SSRN Electronic Journal*.

Haves, E. (2022, December 8). *Regulation and practices of private equity*. Retrieved from House of Lords Library: https://lordslibrary.parliament.uk/regulation-and-practices-of-private-equity/

Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom You Know Matters: Venture Capital. *The Journal of Finance, 62*(1), 251-301.

Investment Company Act of 1940 (August 22, 1940).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning.*

Kaplan, S. N., & Schoar, A. (2005). Private equity performance: Returns, persistence, and capital flows. *Journal of Finance, 60*(4), 1791-1823.

Lerner, J., Leamon, A., & Hardymon, F. (2012). *Venture Capital Private Equity, and the Financing of Entrepreneurship.* John Wiley & Sons, Inc.

Ljungqvist, A., & Richardson, M. P. (2003). The investment behavior of private equity fund managers. *RICAFE Working paper*.

Loos, N. (2006). *Value Creation in Leveraged Buyouts: Analysis of Factors Driving Private Equity Investment Performance.* Wiesbaden, Germany: Wiesbaden : Deutscher Universitätsverlag.

Lossen, U. (2006). The Performance of Private Equity Funds: Does Diversification Matter? *Munich Business Research Working Paper Series No. 2006-14*, 1747–1776.

Manigart, S., De Waele, K., Wright, M., Robbie, K., Desbrières, P., Sapienza, H., & Beekman, A. (2002). Determinants of required return in venture capital investments: A five country study. *Journal of Business Venturing, 17*(4).

McSwain, P. (2019, June 25). Private Equity Access: Should We Beware?

Moonfare GmbH. (2022, December 10). *Private Equity Basics in 6 Charts*. Retrieved from Moonfare GmbH Website: https://www.moonfare.com/blog/private-equity-basics-in-6-charts

MSCI. (2022, December 10). *GICS*. Retrieved from MSCI Website: https://www.msci.com/our-solutions/indexes/gics

Office of Compliance Inspections and Examinations. (2020). *Observations from Examinations of Investment Advisers Managing Private Funds.*

Paul Hastings. (2020, June 9). Department of Labor Allows Private Equity Investment Exposure in 401(k) Plans.

Phalippou, L., & Zollo, M. (2005). What Drives Private Equity Fund Performance? *INSEAD Working paper series*.

PitchBook. (2022a, December 10). *What is the difference between IRR and PME?* Retrieved from PitchBook Website: https://pitchbook.com/blog/what-is-the-difference-between-irr-and-pme

PitchBook. (2022b, December 10). *What is private market benchmarking and why is it important?* Retrieved from PitchBook Website: https://pitchbook.com/blog/what-are-private-market-benchmarks-how-are-they-used-and-why-do-they-matter

PitchBook. (2022c, 8 December). *How PitchBook Collects Data*. Retrieved from PitchBook website: https://pitchbook.com/research-process

Roggi, O., & Giannozzi, A. (2019). Private equity characteristics and performance: An analysis of North American venture capital and buyout funds. *Economic Notes, 48*(2).

Rouvinez, C. (2014). Private Equity Benchmarking with PME. *Private Equity International*.

Securities and Exchange Commission. (2020). *Final Rule: Amending the "Accredited Investor" Definition.*

Sexton, L., & Veronis, N. (2021, February 3). Private Markets Have Become Accessible to Institutional Investors .

Sigrist, N., & Perfetto, M. (2019, October). A Quantitative Approach to Private Equity Fund Selection. *Perspectives*.

The World Bank. (2022). *Listed Domestic Companies, total - United States*. Retrieved from data.Worldbank.org: https://data.worldbank.org/indicator/CM.MKT.LDOM.NO?locations=US

Tiozzo, V., & Morales, M. (2019). *Machine Learning and Survival Prediction Models for Decision Support in the Private Equity Market.*

United Nations. (2022, December 10). *Standard country or area codes for statistical use (M49)*. Retrieved from United Nations Wesite: https://unstats.un.org/unsd/methodology/m49/

United States Census Bureau. (2022a, December 10). *Census Regions and Divisions of the United States*. Retrieved from United States Census Bureau Website: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

United States Census Bureau. (2022b, December 10). *Measuring America's People, Places, and Economy*. Retrieved from United States Census Bureau Website: https://www.census.gov/

Data Sources

This list presents the macroeconomic data which was collected from publicly available sources:

[1] Stock Market Returns:

Investing.com. (2022, November 9). *S&P 500 (SPX)*. Retrieved from Investing.com: https://www.investing.com/indices/us-spx-500-historical-data

Investing.com. (2022, November 9). *STOXX 600 (STOXX)*. Retrieved from Investing.com: https://www.investing.com/indices/stoxx-600-historical-data

Backtest. (2022, November 9). *Historical data of the MSCI Emerging Markets index*. Retrieved from Backtest by Curvo: https://backtest.curvo.eu/market-index/msci-emerging-markets

[2] Bond Yield:

Investing.com. (2022, November 9). *United States 10-Year Bond Yield*. Retrieved from Investing.com: https://www.investing.com/rates-bonds/u.s.-10-year-bond-yield

European Central Bank. (2022, November 9). *Euro area 10-year Government Benchmark bond yield.* Retrieved from European Central Bank Website: https://sdw.ecb.europa.eu/quickview.do?SERIES_KEY=143.FM.M.U2.EUR.4F.BB.U2_10Y.YLD&fbclid=IwAR1CWDY1HxFWvUEHd7QJ3Kndq1ORvk4P7msa2bHbUoegl7s9VSk-4Dxs3DI&

Bloomberg. (2022, November 9). *JPMorgan Emerging Markets Bond Index (EMBI)*. Retrieved from Bloomberg Terminal.

[3] Credit Spread:

Federal Reserve Bank Of St. Louis. (2022, November 9). *ICE BofA US High Yield Index Option-Adjusted Spread*. Retrieved from Federal Reserve Bank Of St. Louis Website: https://fred.stlouisfed.org/series/BAMLH0A0HYM2?fbclid=IwAR2TGl-mZawUolITue7QXjVW-niyh6Whr7zTLTO0g_wHcVi-06gSQYfAPnA

Federal Reserve Bank Of St. Louis. (2022, November 9). *ICE BofA Euro High Yield Index Option-Adjusted Spread*. Retrieved from Federal Reserve Bank Of St. Louis: https://fred.stlouisfed.org/series/BAMLHE00EHYIOAS?fbclid=IwAR31Ke_dOJZaYNBd_ddBP8NelUhs827cKFlyL013iEsTwRLhhin3MJAtyyY

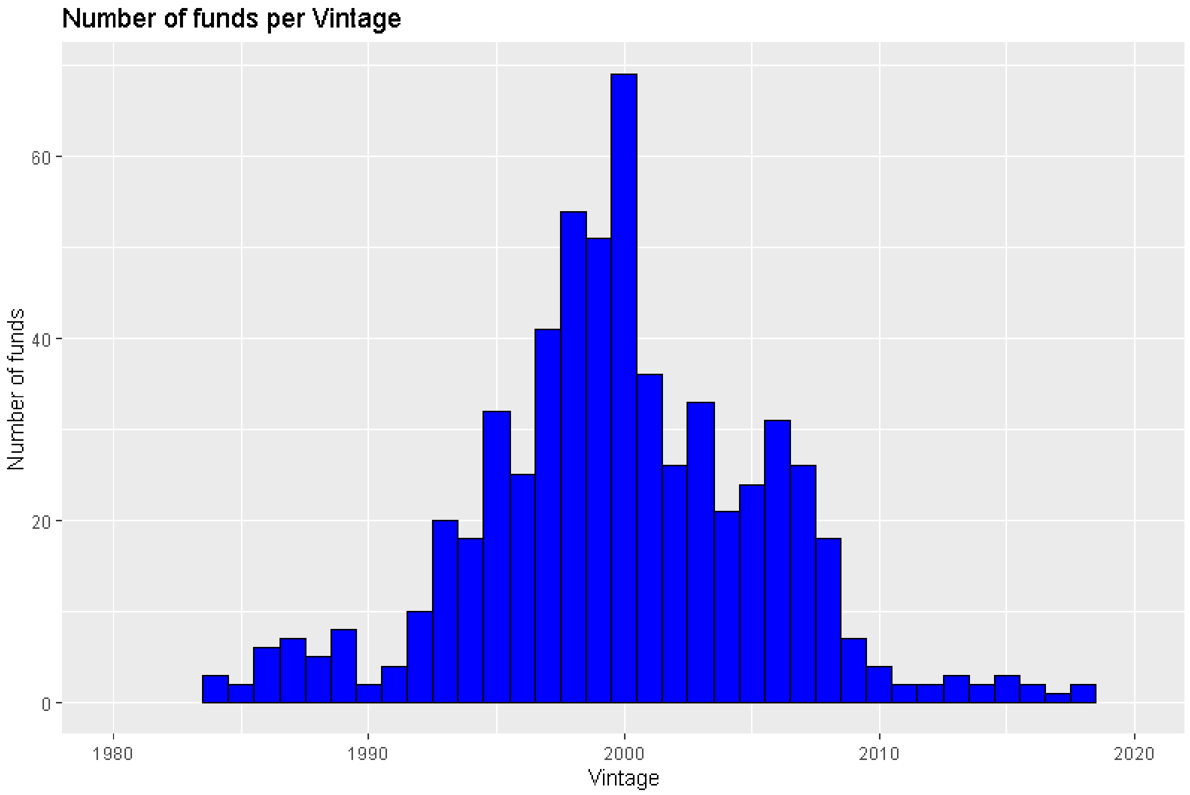
Federal Reserve Bank Of St. Louis. (2022, November 9). *ICE BofA High Yield Emerging Markets Corporate Plus Index Option-Adjusted Spread*. Retrieved from Federal Reserve Bank Of St. Louis: https://fred.stlouisfed.org/series/BAMLEMHBHYCRPIOAS?fbclid=IwAR0kJwQP3HN4B4lMe16mqhPxi_sLX2tiuuhSSJmA31SugOTMbVXFX5OOu2E

[4] GDP Growth:

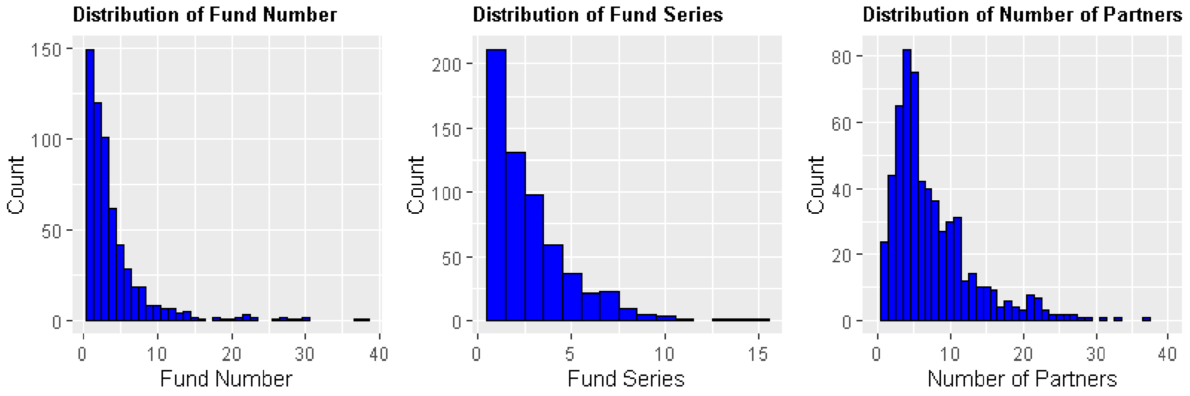
International Monetary Fund. (2022, November 9). *Real GDP growth: Annual percent change*. Retrieved from International Monetary Fund Website: https://www.imf.org/external/datamapper/NGDP_RPCH@WEO/WEOWORLD/OEMDC/EUQ/USA

Appendix A

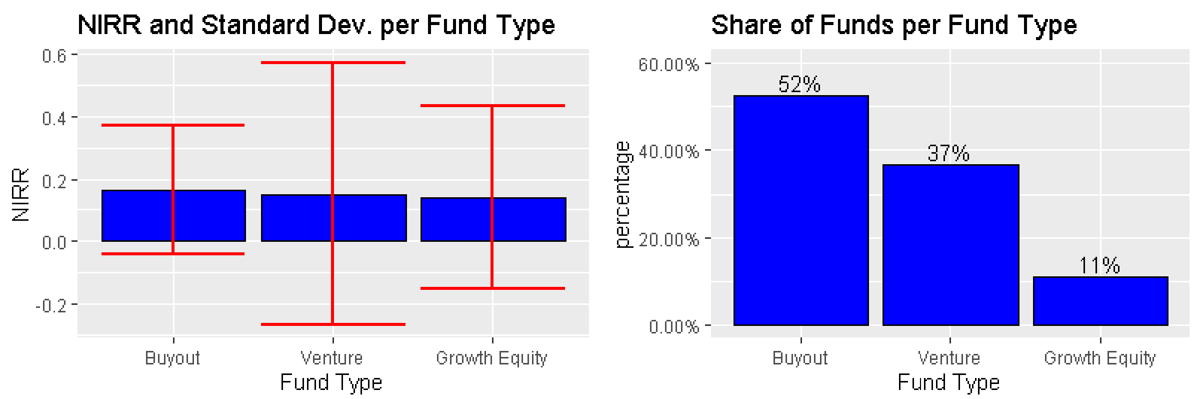
A.1 Variables Distributions



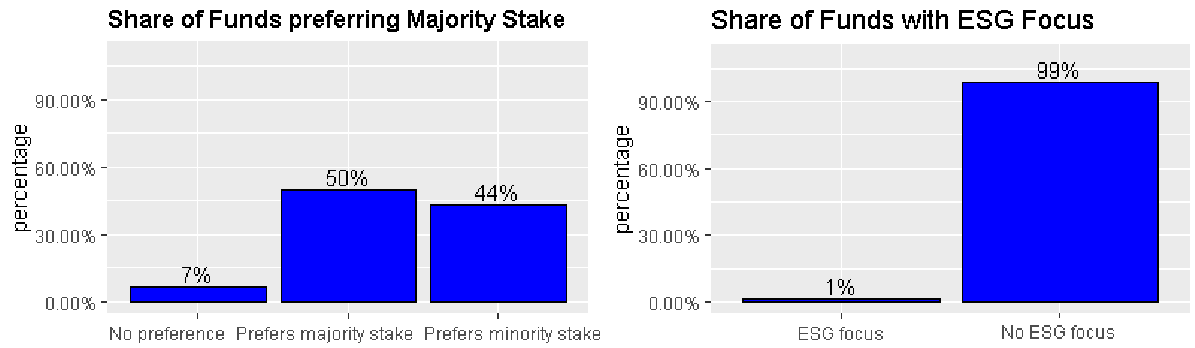
(a) Vintage



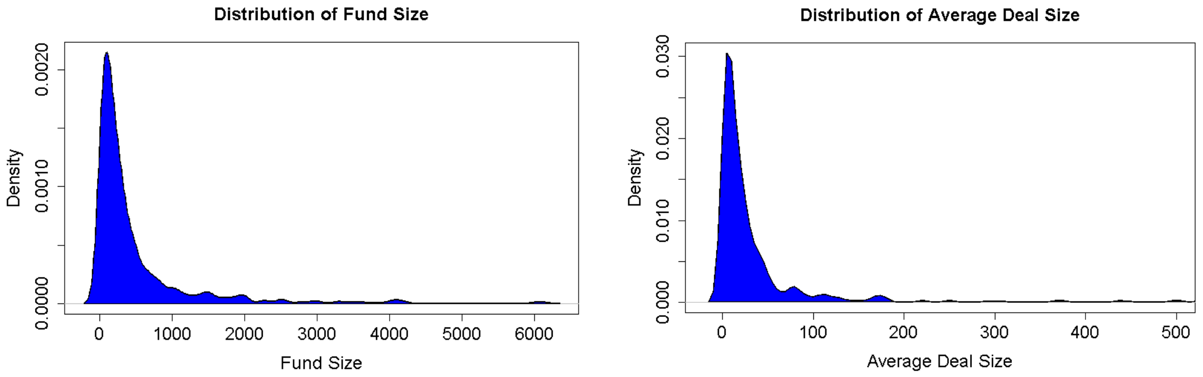
(b) Management Experience



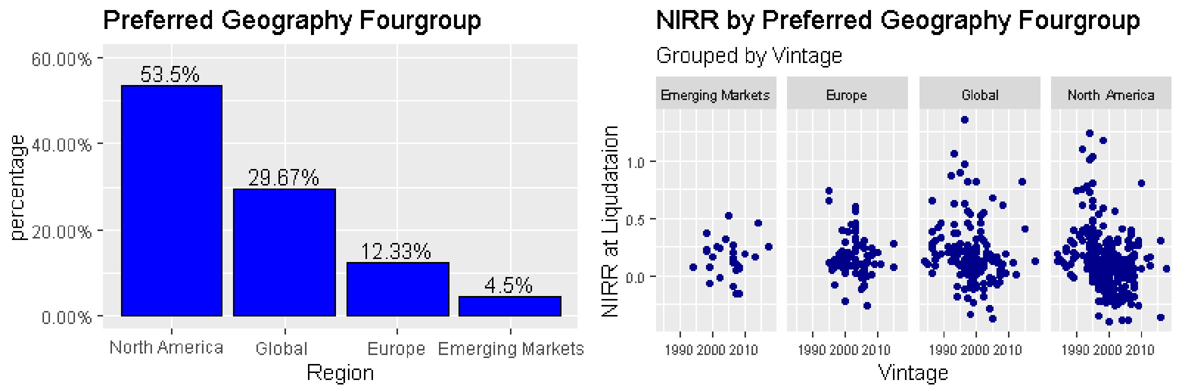
(c) Fund Type



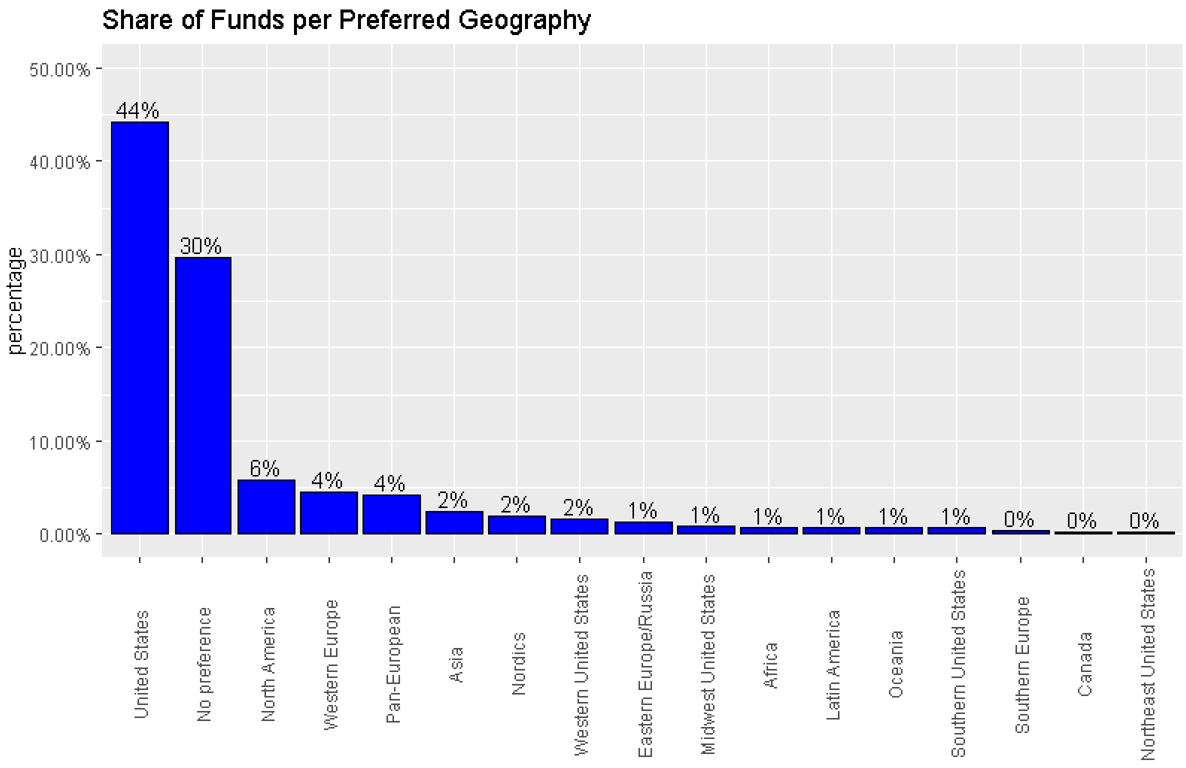
(d) Majority Stake Preference and ESG Focus



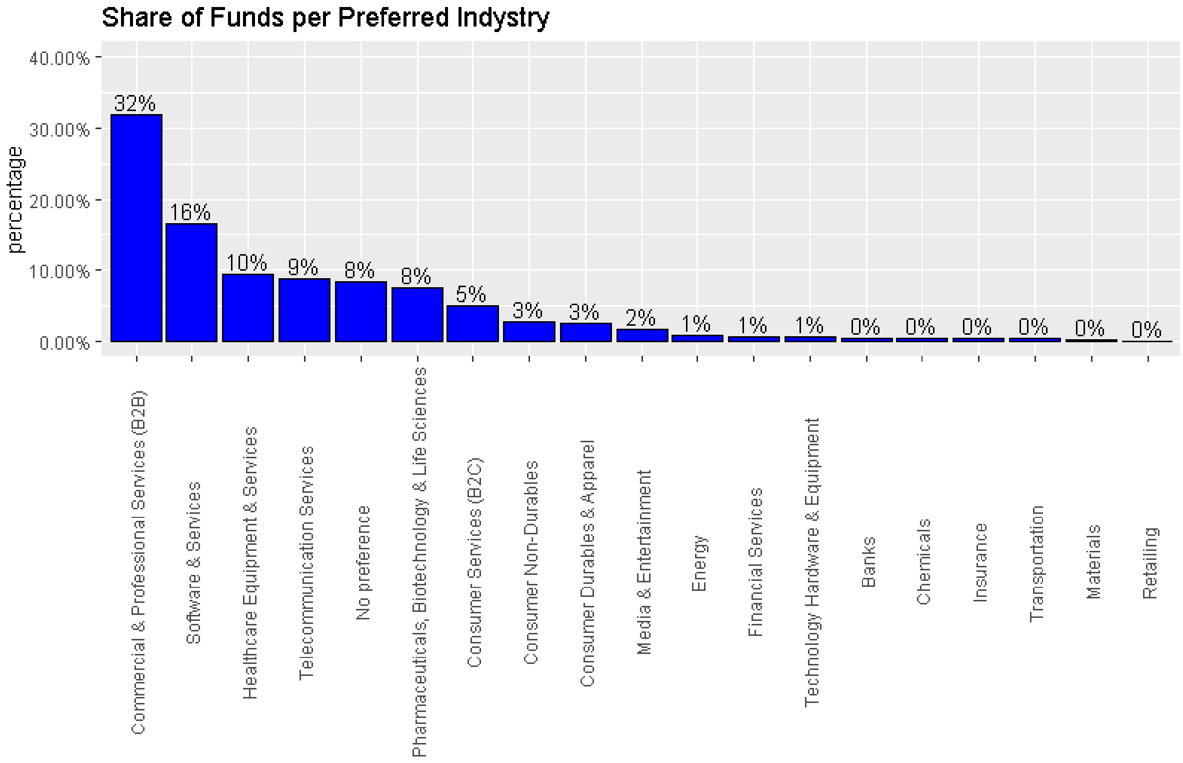
(e) Fund Size and Average Deal Size



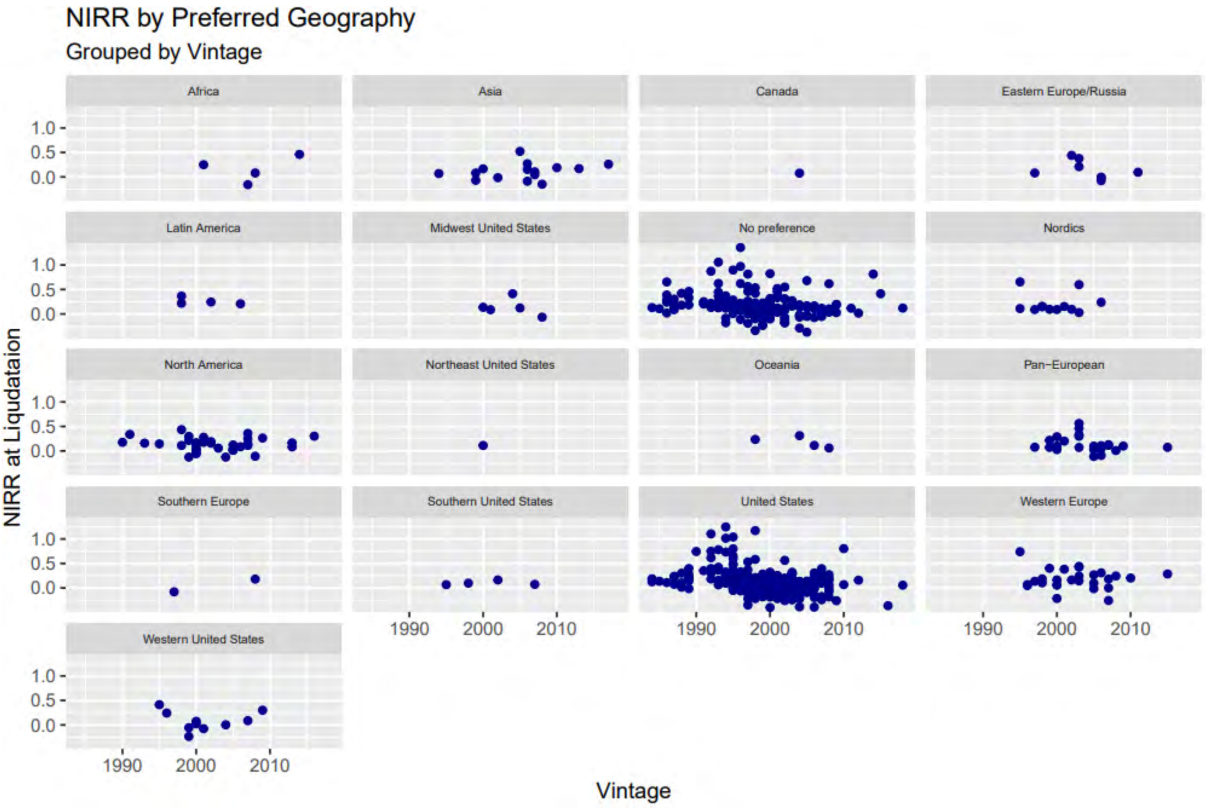
(f) Preferred Geography Fourgroup and NIRR by Preferred Geography Fourgroup



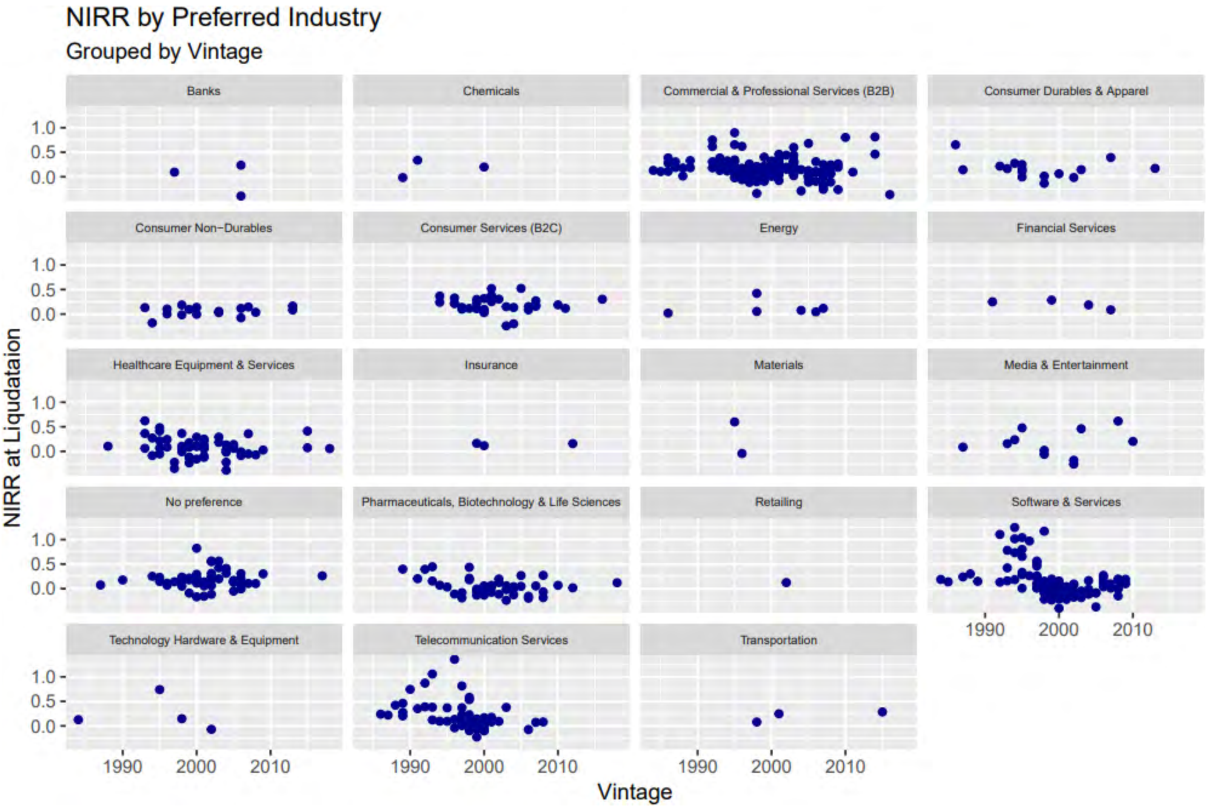
(g) Preferred Geography



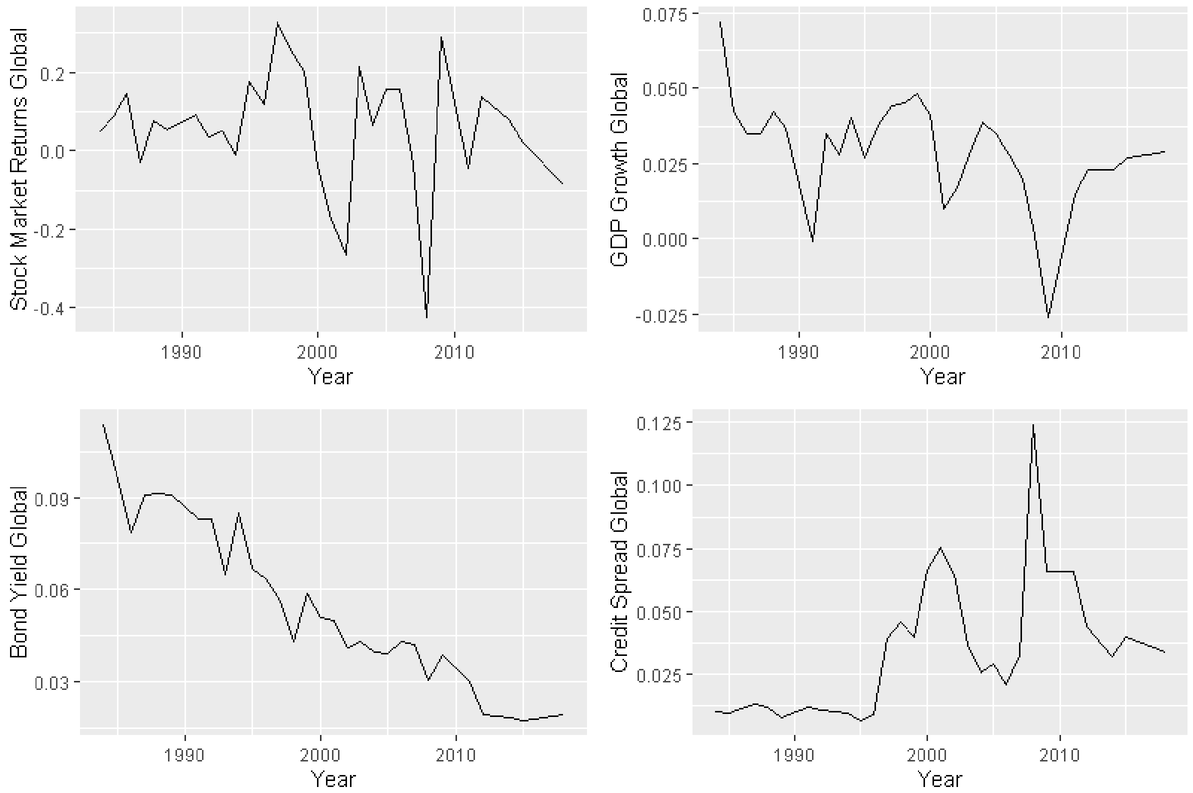
(h) Preferred Industry



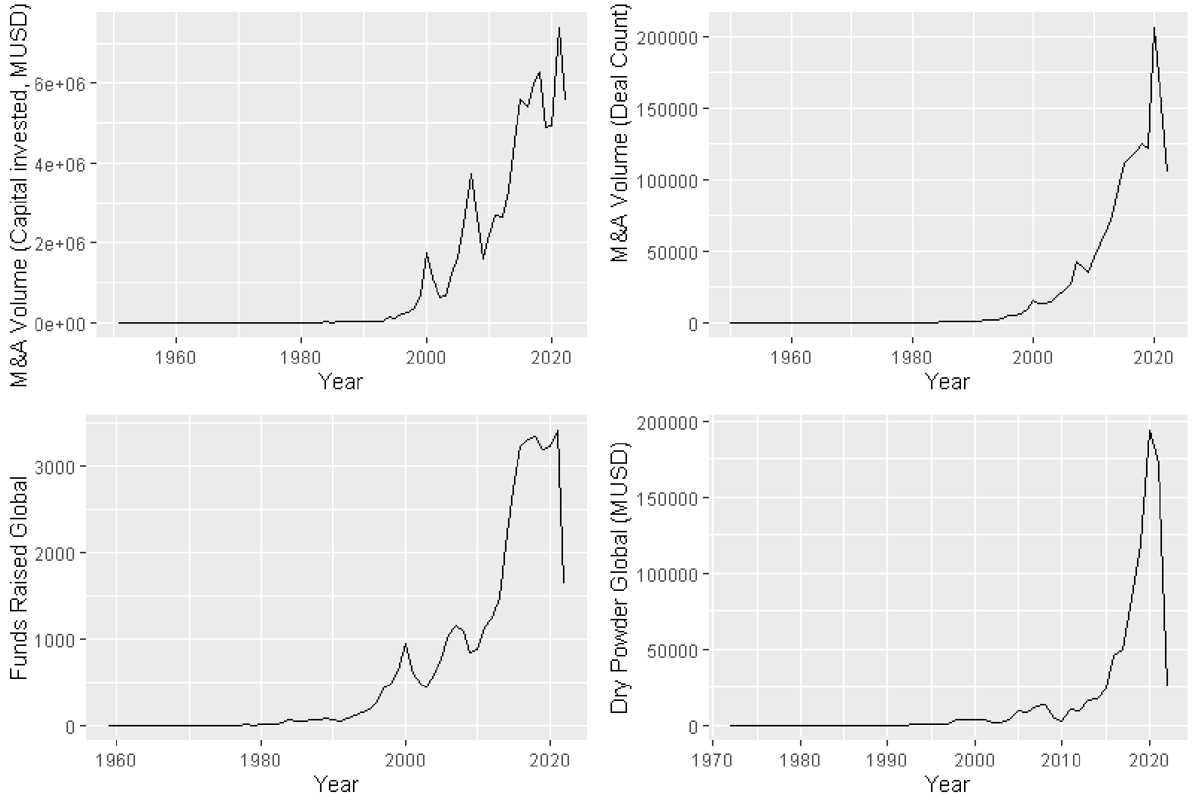
(i) NIRR by Preferred Geography

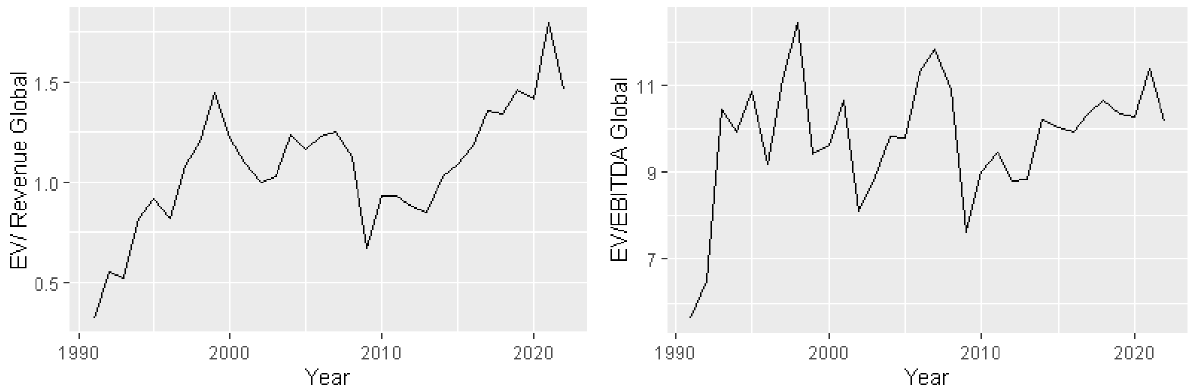


(j) NIRR by Preferred Industry



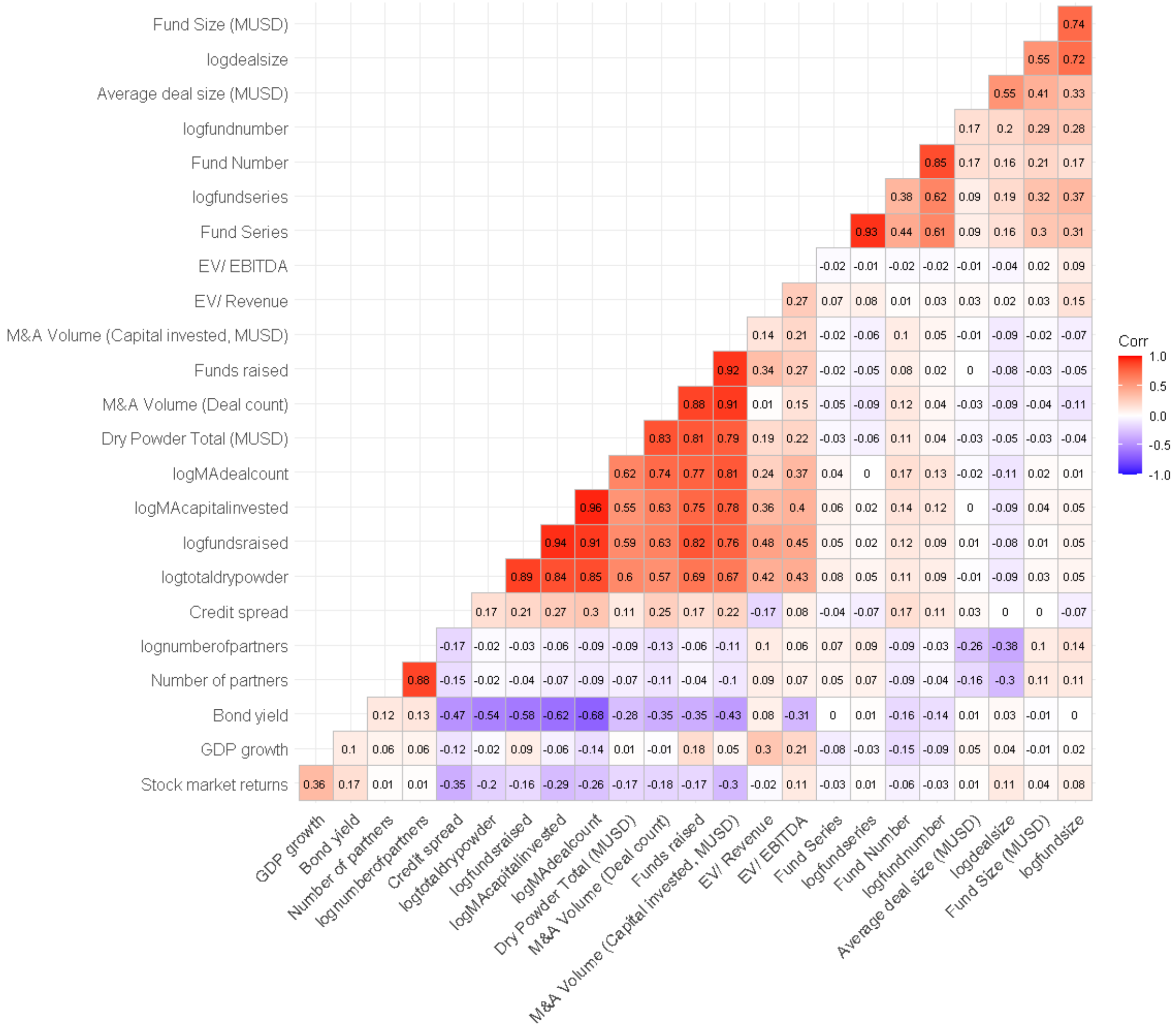
(k) Macroeconomic Variables





(l) PE Market Variables

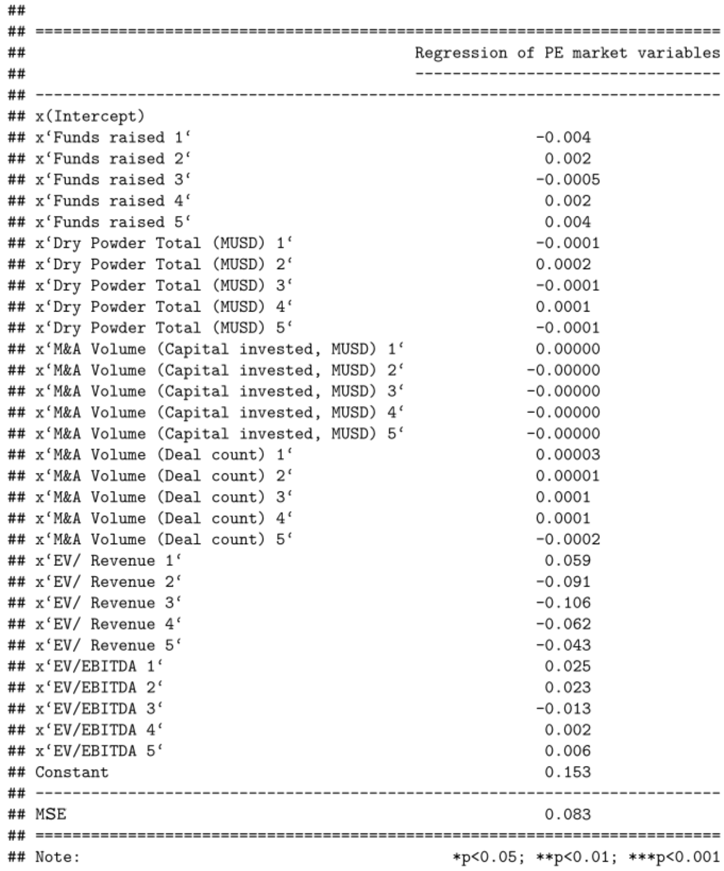
A.2 Variables Correlation Matrix



	GDP growth	Bond yield	Number of partners	lognumberofpartners	Credit spread	logtotaldrypowder	logfundsraised	logMAcapitalinvested	logMAdealcount	Dry Powder Total (MUSD)	M&A Volume (Deal count)	Funds raised	M&A Volume (Capital invested, MUSD)	EV/ Revenue	EV/ EBITDA	Fund Series	logfundseries	Fund Number	logfundnumber	Average deal size (MUSD)	logdealsize	Fund Size (MUSD)	logfundsize
Fund Size (MUSD)																							0.74
logdealsize																						0.55	0.72
Average deal size (MUSD)																					0.55	0.41	0.33
logfundnumber																				0.17	0.2	0.29	0.28
Fund Number																			0.85	0.17	0.16	0.21	0.17
logfundseries																		0.38	0.62	0.09	0.19	0.32	0.37
Fund Series																	0.93	0.44	0.61	0.09	0.16	0.3	0.31
EV/ EBITDA																-0.02	-0.01	-0.02	-0.02	-0.01	-0.04	0.02	0.09
EV/ Revenue															0.27	0.07	0.08	0.01	0.03	0.03	0.02	0.03	0.15
M&A Volume (Capital invested, MUSD)														0.14	0.21	-0.02	-0.06	0.1	0.05	-0.01	-0.09	-0.02	-0.07
Funds raised													0.92	0.34	0.27	-0.02	-0.05	0.08	0.02	0	-0.08	-0.03	-0.05
M&A Volume (Deal count)												0.88	0.91	0.01	0.15	-0.05	-0.09	0.12	0.04	-0.03	-0.09	-0.04	-0.11
Dry Powder Total (MUSD)											0.83	0.81	0.79	0.19	0.22	-0.03	-0.06	0.11	0.04	-0.03	-0.05	-0.03	-0.04
logMAdealcount										0.62	0.74	0.77	0.81	0.24	0.37	0.04	0	0.17	0.13	-0.02	-0.11	0.02	0.01
logMAcapitalinvested									0.96	0.55	0.63	0.75	0.78	0.36	0.4	0.06	0.02	0.14	0.12	0	-0.09	0.04	0.05
logfundsraised								0.94	0.91	0.59	0.63	0.82	0.76	0.48	0.45	0.05	0.02	0.12	0.09	0.01	-0.08	0.01	0.05
logtotaldrypowder							0.89	0.84	0.85	0.6	0.57	0.69	0.67	0.42	0.43	0.08	0.05	0.11	0.09	-0.01	-0.09	0.03	0.05
Credit spread						0.17	0.21	0.27	0.3	0.11	0.25	0.17	0.22	-0.17	0.08	-0.04	-0.07	0.17	0.11	0.03	0	0	-0.07
lognumberofpartners					-0.17	-0.02	-0.03	-0.06	-0.09	-0.09	-0.13	-0.06	-0.11	0.1	0.06	0.07	0.09	-0.09	-0.03	-0.26	-0.38	0.1	0.14
Number of partners				0.88	-0.15	-0.02	-0.04	-0.07	-0.09	-0.07	-0.11	-0.04	-0.1	0.09	0.07	0.05	0.07	-0.09	-0.04	-0.16	-0.3	0.11	0.11
Bond yield			0.12	0.13	-0.47	-0.54	-0.58	-0.62	-0.68	-0.28	-0.35	-0.35	-0.43	0.08	-0.31	0	0.01	-0.16	-0.14	0.01	0.03	-0.01	0
GDP growth		0.1	0.06	0.06	-0.12	-0.02	0.09	-0.06	-0.14	0.01	-0.01	0.18	0.05	0.3	0.21	-0.08	-0.03	-0.15	-0.09	0.05	0.04	-0.01	0.02
Stock market returns	0.36	0.17	0.01	0.01	-0.35	-0.2	-0.16	-0.29	-0.26	-0.17	-0.18	-0.17	-0.3	-0.02	0.11	-0.03	0.01	-0.06	-0.03	0.01	0.11	0.04	0.08

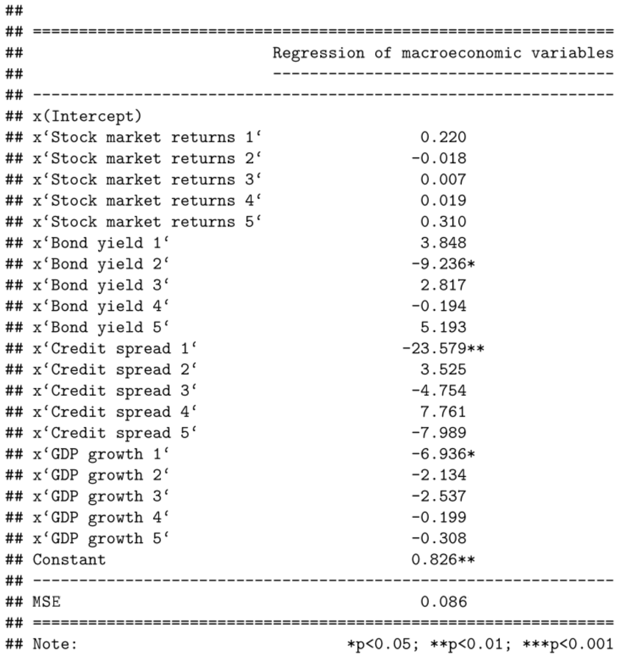
Appendix B

B.1 Regressions of Macroeconomic & PE Market Variables on NIRR



```
##
## =========================================================================
##                                     Regression of PE market variables
##                                     ---------------------------------
## -------------------------------------------------------------------------
## x(Intercept)
## x‘Funds raised 1‘                                   -0.004
## x‘Funds raised 2‘                                    0.002
## x‘Funds raised 3‘                                   -0.0005
## x‘Funds raised 4‘                                    0.002
## x‘Funds raised 5‘                                    0.004
## x‘Dry Powder Total (MUSD) 1‘                        -0.0001
## x‘Dry Powder Total (MUSD) 2‘                        0.0002
## x‘Dry Powder Total (MUSD) 3‘                        -0.0001
## x‘Dry Powder Total (MUSD) 4‘                        0.0001
## x‘Dry Powder Total (MUSD) 5‘                        -0.0001
## x‘M&A Volume (Capital invested, MUSD) 1‘            0.00000
## x‘M&A Volume (Capital invested, MUSD) 2‘           -0.00000
## x‘M&A Volume (Capital invested, MUSD) 3‘           -0.00000
## x‘M&A Volume (Capital invested, MUSD) 4‘           -0.00000
## x‘M&A Volume (Capital invested, MUSD) 5‘           -0.00000
## x‘M&A Volume (Deal count) 1‘                        0.00003
## x‘M&A Volume (Deal count) 2‘                        0.00001
## x‘M&A Volume (Deal count) 3‘                        0.0001
## x‘M&A Volume (Deal count) 4‘                        0.0001
## x‘M&A Volume (Deal count) 5‘                        -0.0002
## x‘EV/ Revenue 1‘                                     0.059
## x‘EV/ Revenue 2‘                                    -0.091
## x‘EV/ Revenue 3‘                                    -0.106
## x‘EV/ Revenue 4‘                                    -0.062
## x‘EV/ Revenue 5‘                                    -0.043
## x‘EV/EBITDA 1‘                                       0.025
## x‘EV/EBITDA 2‘                                       0.023
## x‘EV/EBITDA 3‘                                      -0.013
## x‘EV/EBITDA 4‘                                       0.002
## x‘EV/EBITDA 5‘                                       0.006
## Constant                                             0.153
## -------------------------------------------------------------------------
## MSE                                                  0.083
## =========================================================================
## Note:                                     *p<0.05; **p<0.01; ***p<0.001
```

(a) PE Market Variables Regressed on NIRR



```
##
## ==============================================================
##                             Regression of macroeconomic variables
##                             -------------------------------------
## --------------------------------------------------------------
## x(Intercept)
## x‘Stock market returns 1‘              0.220
## x‘Stock market returns 2‘             -0.018
## x‘Stock market returns 3‘              0.007
## x‘Stock market returns 4‘              0.019
## x‘Stock market returns 5‘              0.310
## x‘Bond yield 1‘                        3.848
## x‘Bond yield 2‘                       -9.236*
## x‘Bond yield 3‘                        2.817
## x‘Bond yield 4‘                       -0.194
## x‘Bond yield 5‘                        5.193
## x‘Credit spread 1‘                   -23.579**
## x‘Credit spread 2‘                     3.525
## x‘Credit spread 3‘                    -4.754
## x‘Credit spread 4‘                     7.761
## x‘Credit spread 5‘                    -7.989
## x‘GDP growth 1‘                       -6.936*
## x‘GDP growth 2‘                       -2.134
## x‘GDP growth 3‘                       -2.537
## x‘GDP growth 4‘                       -0.199
## x‘GDP growth 5‘                       -0.308
## Constant                               0.826**
## --------------------------------------------------------------
## MSE                                    0.086
## ==============================================================
## Note:                             *p<0.05; **p<0.01; ***p<0.001
```

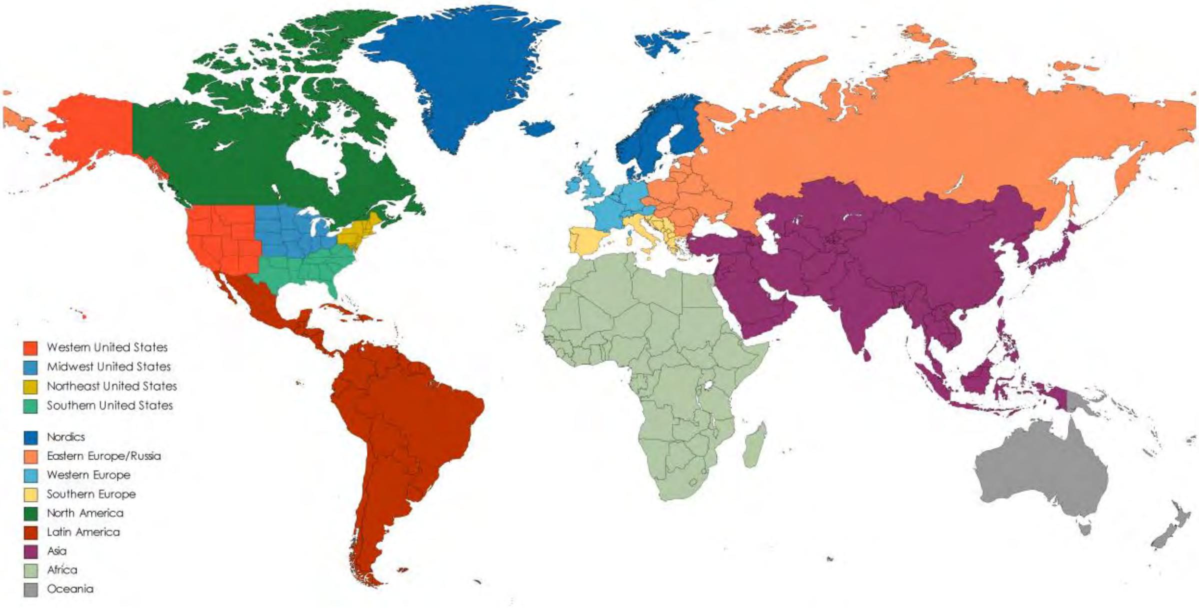
(b) Macroeconomic Variables Regressed on NIRR

B.2 Industry & Geographic Region Classification

Sector	Industry Group	Industry
Energy	**Energy**	Energy Equipment & Services
		Oil, Gas & Consumable Fuels
Materials	**Materials**	Chemicals
		Construction Chemicals
		Container & Packaging
		Metals & Mining
		Paper & Forest Products
Industrials	**Capital Goods**	Aerospace & Defence
		Building Product
		Construction & Engineering
		Electrical Equipment
		Industrial Conglomerates
		Machinery
		Trading Companies & Distributors
	Commercial & Professional Services	Commercial Services & Supplies
		Professional Services
	Transportation	Air Fright & Logistics
		Airlines
		Marine
		Road & Rail
		Transportation Infrastructure
Consumer Discretionary (Consumer Cyclical)	**Automobiles & Components**	Auto Components
		Automobiles
	Consumer Durables & Apparel	Household Durables
		Leisure Products
		Textiles, Apparel & Luxury Goods
	Consumer Services	Hotels, Restaurants & Leisure
		Diversified Consumer Services
	Retailing	Distributors
		Internet & Direct Marketing Retail
		Multiline Retail
		Speciality Retail
Consumer Staples (Consumer Defensive)	**Food & Staples Retailing**	Food & Staples Retailing
	Food, Beverage & Tobacco	Beverages
		Food Products
		Tobacco
	Household & Personal Products	Household Products
		Personal Products

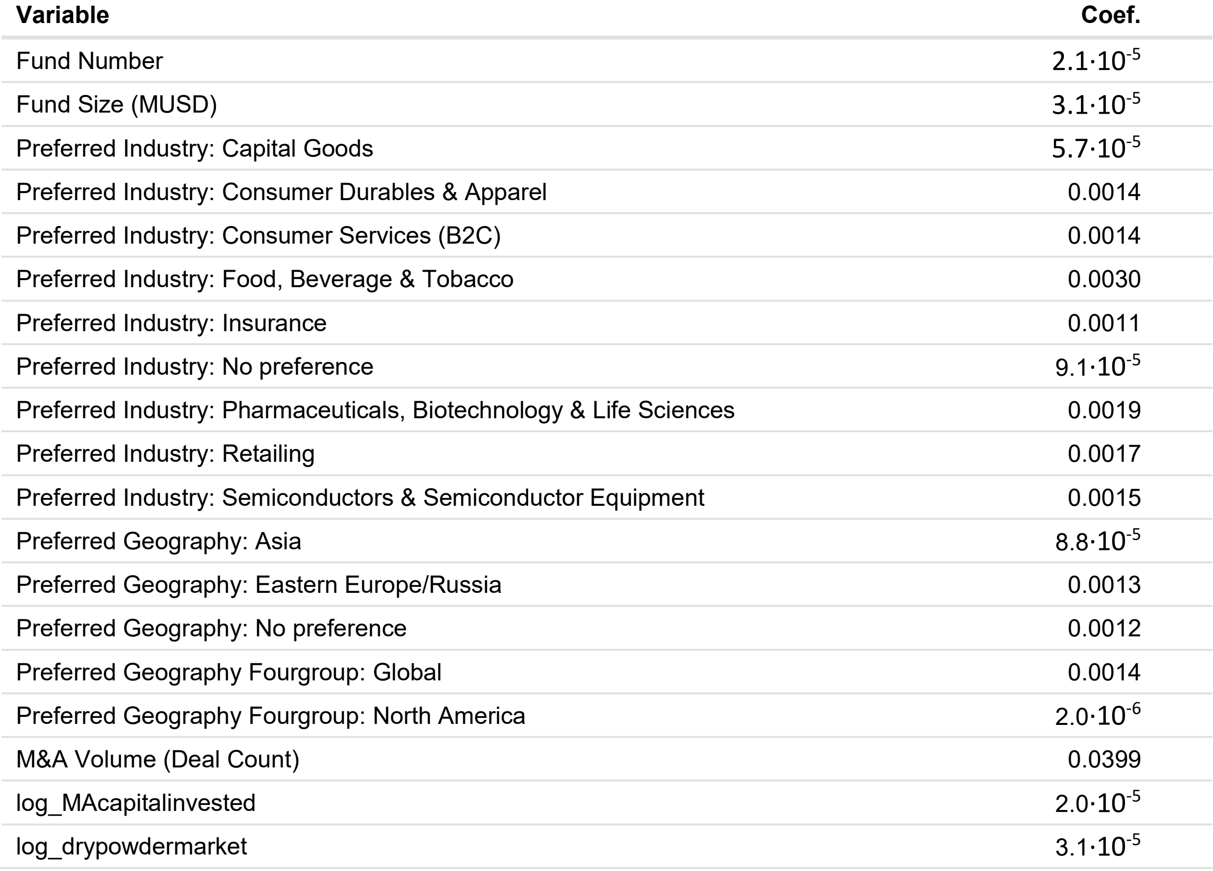
Healthcare	**Healthcare Equipment & Services**	Health Care Equipment & Supplies
		Health Care Providers & Services
		Health Care Technology
	Pharmaceuticals, Biotechnology & Life Sciences	Biotechnology
		Pharmaceuticals
		Life Sciences Tools & Services
Financials	**Banks**	Banks
		Thrifts & Mortgage Finance
	Financial Services	Diversified Financial Services
		Consumer Finance
		Capital Markets
		Mortgage Real Estate Investment Trusts (REITs)
	Insurance	Insurance
Information Technology	**Software & Services**	IT Services
		Software
	Technology Hardware & Equipment	Communications Equipment
		Technology Hardware, Storage & Peripherals
		Electronic Equipment, Instruments & Components
	Semiconductors & Semiconductor Equipment	Semiconductors & Semiconductor Equipment
Communication Services	**Telecommunication Services**	Diversified Telecommunication Services
		Wireless Telecommunication Services
	Media & Entertainment	Media
		Entertainment
		Interactive Media & Services
Utilities	**Utilities**	Electric Utilities
		Gas Utilities
		Multi-Utilities
		Water Utilities
		Independent Power and Renewable Electricity Producers
Real Estate	**Real Estate**	Equity Real Estate Investment Trusts (REITs)
		Real Estate Management & Development

(a) Global Industry Classification Standard (GICS) for Preferred Industry Variable



(b) Geographic Regions Classification for Preferred Geography Variable

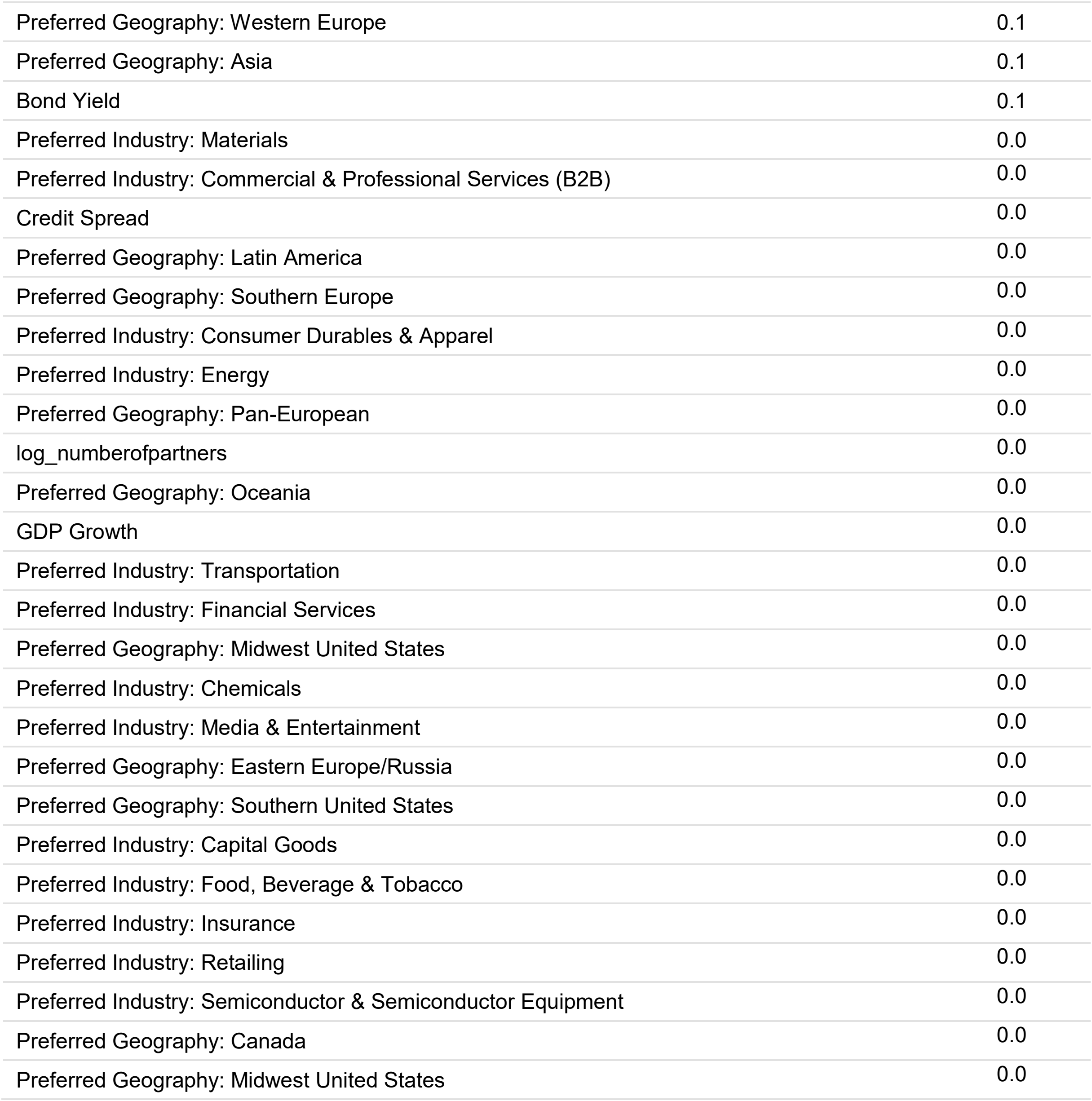
B.3 ML Models Coefficients Complete Overview



Variable	Coef.
Fund Number	$2.1\cdot10^{-5}$
Fund Size (MUSD)	$3.1\cdot10^{-5}$
Preferred Industry: Capital Goods	$5.7\cdot10^{-5}$
Preferred Industry: Consumer Durables & Apparel	0.0014
Preferred Industry: Consumer Services (B2C)	0.0014
Preferred Industry: Food, Beverage & Tobacco	0.0030
Preferred Industry: Insurance	0.0011
Preferred Industry: No preference	$9.1\cdot10^{-5}$
Preferred Industry: Pharmaceuticals, Biotechnology & Life Sciences	0.0019
Preferred Industry: Retailing	0.0017
Preferred Industry: Semiconductors & Semiconductor Equipment	0.0015
Preferred Geography: Asia	$8.8\cdot10^{-5}$
Preferred Geography: Eastern Europe/Russia	0.0013
Preferred Geography: No preference	0.0012
Preferred Geography Fourgroup: Global	0.0014
Preferred Geography Fourgroup: North America	$2.0\cdot10^{-6}$
M&A Volume (Deal Count)	0.0399
log_MAcapitalinvested	$2.0\cdot10^{-5}$
log_drypowdermarket	$3.1\cdot10^{-5}$

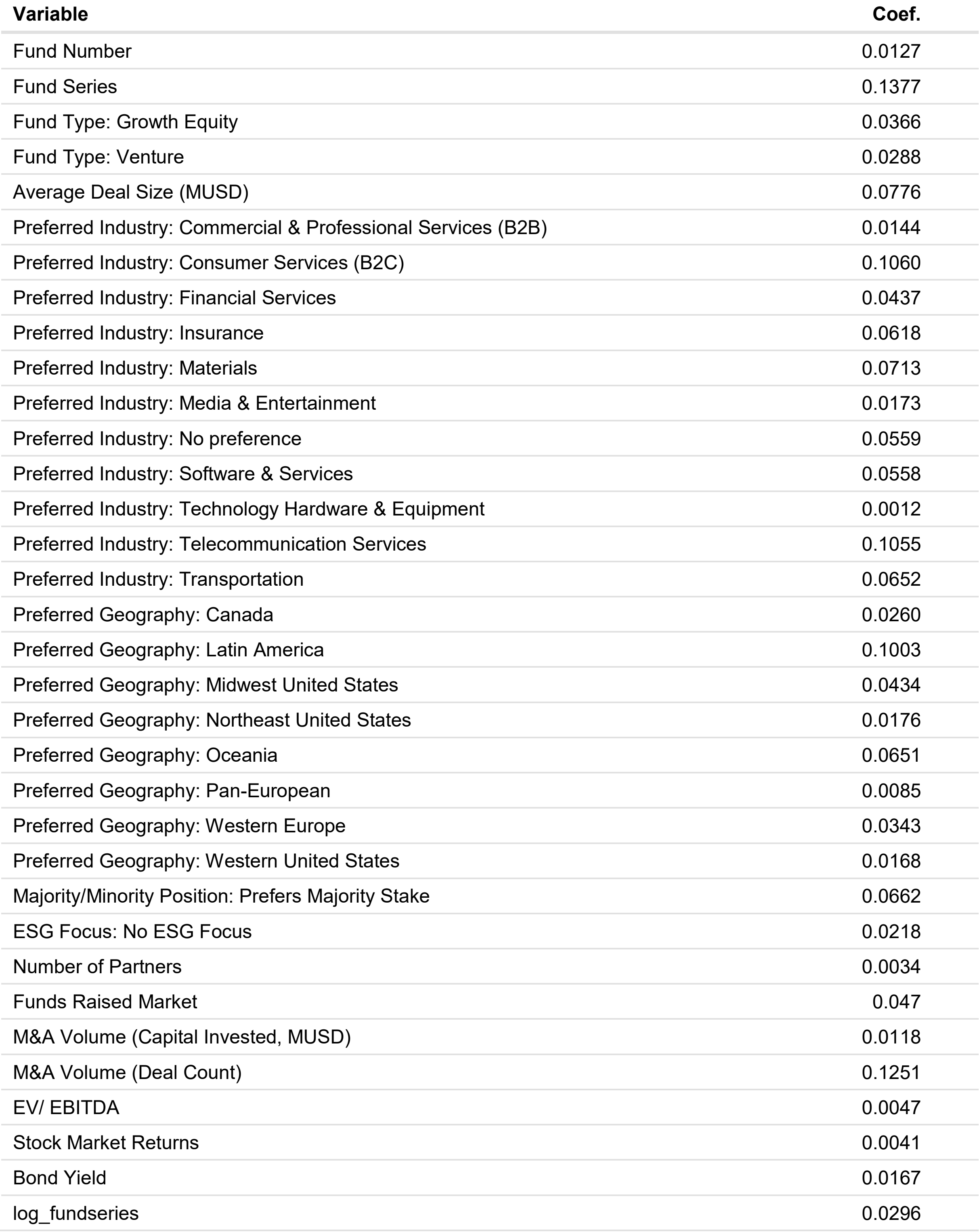
(a) Coefficients of Ridge Model

Variable	Squared coef.
M&A Volume (Capital Invested, MUSD)	$1.2 \cdot 10^{14}$
M&A Volume (Deal Count)	$9.5 \cdot 10^{9}$
Dry Powder Market (MUSD)	$3.3 \cdot 10^{9}$
Funds Raised Market	$2.3 \cdot 10^{7}$
Fund Size (MUSD)	$6.6 \cdot 10^{6}$
Average Deal Size (MUSD)	$4.0 \cdot 10^{4}$
log_M&Acapitalinvested	1,691
log_drypowdermarket	1,614
EV/Revenue	1,057
EV/EBITDA	1,013
log_M&Adealcount	739
log_fundsraisedtotal	521
Number of Partners	152
Fund Series	112
log_fundsize	86
log_averagedealsize	53
Preferred Geography: No preference	15
Preferred Industry: Pharmaceuticals, Biotechnology & Life Sciences	14
Preferred Industry: Healthcare Equipment & Services	11
Preferred Industry: Telecommunication Services	9
Fund Number	7
Majority/Minority Stake: Prefers Majority Stake	5
log_fundnumber	5
Preferred Industry: Software & Services	4
log_fundseries	4
Preferred Geography Fourgroup: North America	4
Preferred Geography Fourgroup: Global	4
Preferred Geography: United States	3
Fund Type: Venture	3
Majority/Minority Stake: Prefers Minority Stake	2
Preferred Geography: Europe	1
Preferred Industry: No Preference	1
Stock Market Returns	1
Preferred Industry: Consumer Non-Durables	0.5
Preferred Geography: Nordics	0.5
Preferred Industry: Consumer Services (B2C)	0.4
Fund Type: Growth Equity	0.4
Preferred Geography: North America	0.3
Preferred Geography: Western United States	0.2
ESG Focus: No ESG Focus	0.2
Preferred Industry: Technology Hardware & Equipment	0.1



Preferred Geography: Western Europe	0.1
Preferred Geography: Asia	0.1
Bond Yield	0.1
Preferred Industry: Materials	0.0
Preferred Industry: Commercial & Professional Services (B2B)	0.0
Credit Spread	0.0
Preferred Geography: Latin America	0.0
Preferred Geography: Southern Europe	0.0
Preferred Industry: Consumer Durables & Apparel	0.0
Preferred Industry: Energy	0.0
Preferred Geography: Pan-European	0.0
log_numberofpartners	0.0
Preferred Geography: Oceania	0.0
GDP Growth	0.0
Preferred Industry: Transportation	0.0
Preferred Industry: Financial Services	0.0
Preferred Geography: Midwest United States	0.0
Preferred Industry: Chemicals	0.0
Preferred Industry: Media & Entertainment	0.0
Preferred Geography: Eastern Europe/Russia	0.0
Preferred Geography: Southern United States	0.0
Preferred Industry: Capital Goods	0.0
Preferred Industry: Food, Beverage & Tobacco	0.0
Preferred Industry: Insurance	0.0
Preferred Industry: Retailing	0.0
Preferred Industry: Semiconductor & Semiconductor Equipment	0.0
Preferred Geography: Canada	0.0
Preferred Geography: Midwest United States	0.0

(b) Squared Coefficients of SVM Model



Variable	Coef.
Fund Number	0.0127
Fund Series	0.1377
Fund Type: Growth Equity	0.0366
Fund Type: Venture	0.0288
Average Deal Size (MUSD)	0.0776
Preferred Industry: Commercial & Professional Services (B2B)	0.0144
Preferred Industry: Consumer Services (B2C)	0.1060
Preferred Industry: Financial Services	0.0437
Preferred Industry: Insurance	0.0618
Preferred Industry: Materials	0.0713
Preferred Industry: Media & Entertainment	0.0173
Preferred Industry: No preference	0.0559
Preferred Industry: Software & Services	0.0558
Preferred Industry: Technology Hardware & Equipment	0.0012
Preferred Industry: Telecommunication Services	0.1055
Preferred Industry: Transportation	0.0652
Preferred Geography: Canada	0.0260
Preferred Geography: Latin America	0.1003
Preferred Geography: Midwest United States	0.0434
Preferred Geography: Northeast United States	0.0176
Preferred Geography: Oceania	0.0651
Preferred Geography: Pan-European	0.0085
Preferred Geography: Western Europe	0.0343
Preferred Geography: Western United States	0.0168
Majority/Minority Position: Prefers Majority Stake	0.0662
ESG Focus: No ESG Focus	0.0218
Number of Partners	0.0034
Funds Raised Market	0.047
M&A Volume (Capital Invested, MUSD)	0.0118
M&A Volume (Deal Count)	0.1251
EV/ EBITDA	0.0047
Stock Market Returns	0.0041
Bond Yield	0.0167
log_fundseries	0.0296

(c) Coefficients of NN Model